Logistics and freight patterns, constraints to productivity and network mapping

Dr Johan W. Joubert

Introduction:

It is rare that people and goods movement in urban areas are considered simultaneously. In this chapter, however, it is argued that they are much more closely related than what we often would like to believe, or even appreciate. The one actually helps us understand the other. In fact, it is a perpetuating cycle. An increasing number of people move to urban areas where there are more employment opportunities. The employers of these people are businesses that provide goods and services to other businesses, and ultimately to the ever-increasing number of consumers in urban areas.

This report is structured into three main themes. In the first we establish the context for this study. That is, we look at the distribution of people in urban areas and highlight the plight of the *poor on the periphery.*

Secondly, we consider the people's need to travel, and mainly focus on trips to work and education opportunities. But in the absence of good data related to places of employment, we introduce commercial vehicle movement as a good proxy. More specifically, we show that where commercial vehicles do their business—pickups and deliveries—are good indications of economic activity, and hence also a good approximation for places of employment. This chapter also introduces a valuable new metric to evaluate the economic accessibility, or *closeness* if you will, of urban areas for doing business.

Finally, we turn back to the people's mobility and look at what variables influence their mode choice. Instead of classic discrete choice models, this chapter introduces Bayesian networks and demonstrates the causal relationships found in the latest National Household Travel Survey. We argue that the models are more intuitive to interpret, making them easier and more useful for decision-support, and also overcome challenges—like incorrect model specification—that burdens the more classical methods.

1. Poor on the periphery

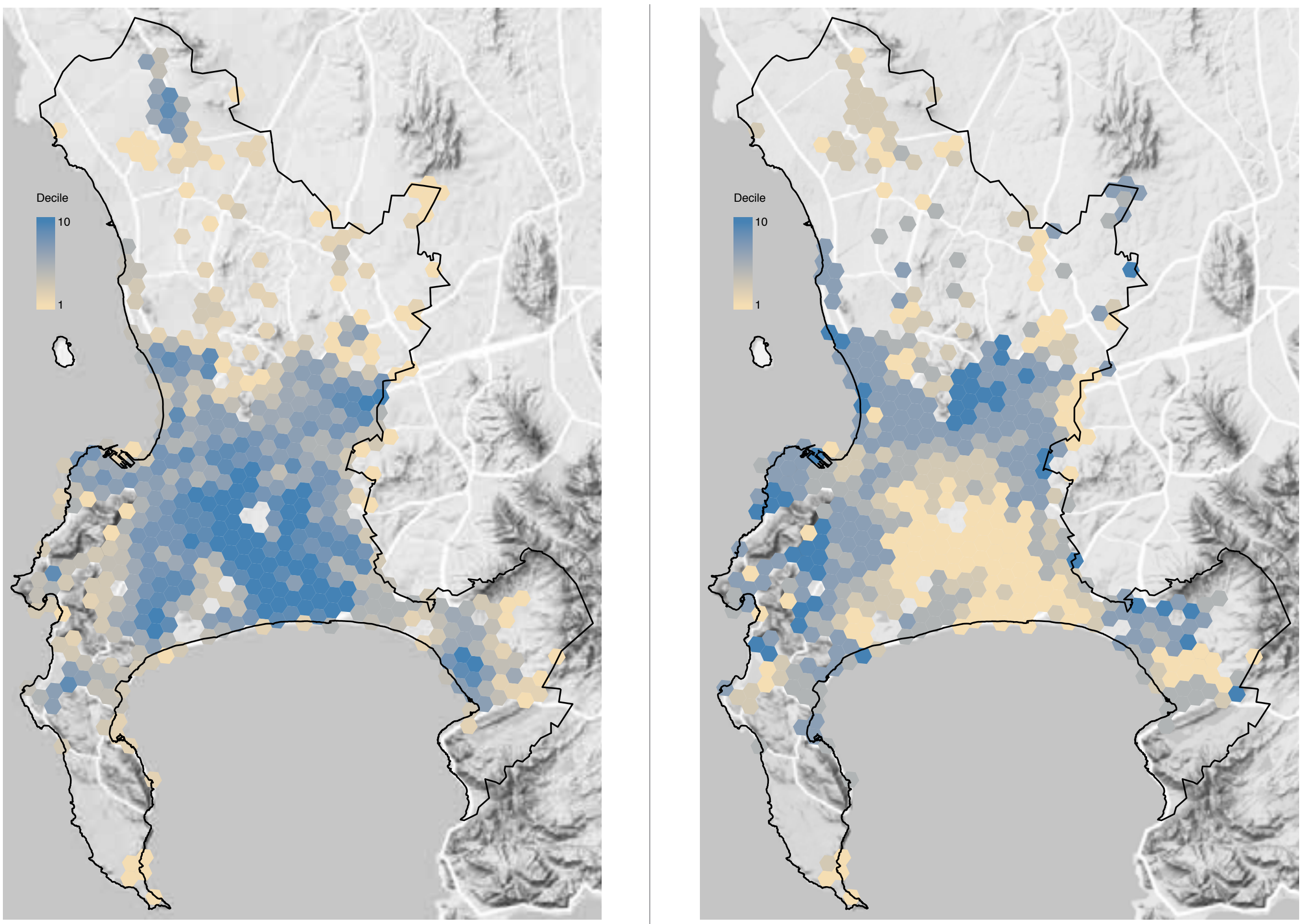
South Africa is well known for its apartheid policies that lead to the majority of non-white citizens being relocated to the periphery of urban areas. Unfortunately, even after twenty years into the country's democracy, little has changed in the urban form. On the contrary, the democratically elected government has reinforced the bad urban form by building more permanent housing in even more peripheral areas where land was cheap. This was an unintended consequence under the Reconstruction and Development Programme (RDP).

A main problem now is that providing quality basic services to these peripheral areas is fiscal suicide for local governments. While Blaise (2011) makes a strong and valid case that the pricing of basic services can indeed curb sprawl when used effectively, it poses some practical challenges in South Africa. Many of the poor on the periphery actually do not pay for basic services at all, and would be economically burdened even more if pricing were used to affect much needed land use change.

In a free market and capitalist environment, economic opportunities often follow the people with buying power. We therefore see the expected economic development happening in areas where good infrastructure and access to economic opportunities already abound.

To put this into context more visually, consider Figure 1 where we compare the population density and the household income in the City of Cape Town (one of the three focus areas of this study). The results ring a too familiar tone. In Figure 1(a) we note the high density of people (dark blue) living in what is commonly known as the Cape Flats. This is also the area we observe in Figure 1b to be in the lowest income decile (predominantly light wheat colour). The results are similar for the eThekwini metro in KwaZulu-Natal, and Gauteng, for which the images are shown in Appendix B. The poor reside on the periphery of economic centra.

Figure 1: Comparing the City of Cape Town's population densities (a) and household incomes (b).



Methodology The density maps are based on a state-of-the-art synthetic population in the public domain (Joubert, 2018). The data is based on the 2011 national census, and benefitted from leading Bayesian network methods published by Sun and Erath (2015) to generate a synthetic population that is accurate at both household and individual level. The population accounts for each member of each household, and includes characteristics such as age, gender, employment/school-going status, household income and dwelling type, the latter being a potential proxy for household income. We choose to aggregate the household data to equal-area zones, and also use deciles as a more descriptive and robust statistic. It allows us to infer the true densities more accurately. Within each zone—a hexagonal area with maximum diameter of 2km—we calculate the median income of all households located in that zone.

As urbanisation increases, many economic hopefuls move to the cities in search of better economic opportunities. While urban densification is generally promoted, very few people consider the resultant effect of freight and commercial vehicles. When urban areas densify, the consumption per area goes up. This increasing consumption pattern must be serviced by (unwanted) commercial vehicles.

Commercial vehicle movement acts as a proxy for economic activity and employment. Trucks and Light Delivery Vehicles (LDVs) perform their activities at the producers, shippers, carriers and receivers of goods. LDVs, especially, is also heavily used in providing service support. These organisations, in turn, do their business by employing people. In turn, people have to use transport to access these activity locations as places of employment, school, shopping, or leisure, to name but a few.

2. Estimating travel demand

In this section we follow a state-of-practice approach to estimate the number of trips in urban areas. That is, the people's demand for travelling for their different purposes. One main aim was to only use existing and publicly available data to do so. Why? To demonstrate that costly (and often non-transparent) studies need not be the norm for authorities when they wish to gain an understanding of travel demand.

2.1 Land-cover data

This study benefits from the *2013–2014 South African National Land-cover Dataset*, a commercial data product by GeoTerraImage that is made available under an open data license to the South African Department of Environmental Affairs. The open data license allows the Department unrestricted access to use the data, and distribute the data unrestricted to third parties.

The 72 class land-cover dataset is based on 30m × 30m raster cells that were derived from satellite imagery and spectral modelling. Each cell is represented as a single pixel in the GeoTIFF file. Associated with each raster cell is a single code representing the dominant land-cover class within the cell[1]. The data set has a reported mean land-cover/land-use class accuracy of 91.27%. From the original 72 classes we focused mainly on classes 42–72 (31 in total) that dealt with built-up areas. The commercial and industrial areas did unfortunately have no subcategories, while residential had numerous subcategories distinguishing between formal, informal, smallholding, township and village types, along with subdivisions indicating whether the surface area is dominated by tree, bush, grass or bare surface.

2.2 Estimating trips

Focusing on built-up areas, we estimated the number of trips generated from and attracted to each raster cell for both the morning and afternoon peaks. The trip estimates are based on the latest *Trip Data Manual* (Committee of Transport Officials, 2013). The manual indicates the Annual Average Daily Trip generation rate (AADT) per area size unit, depending on the land use type. The manual provides generation rates that account for adjustment factors related to peak spreading, percentage of heavy vehicles, and the size of specific developments.

Since the land use types in the trip data manual is broken down into more detail, we inferred weighted averages to estimate the trips from the coarser land cover data set. The resolution of the land-cover dataset precludes the identification of actual buildings, so we prefer to not account for absolute number of trips, but aggregate into deciles over the same equal-area hexagonal zones we introduced with the population and household incomes (Figure 1).

Example: Consider the residential land use for single dwelling units that have an AADT of 4.0 trips per 100m2. Accounting for the variation over the day, the peak-hour generation rate is 1.0 trip/100m2 for the morning (AM) peak, of which the in/out split is 25:75. That is, a raster cell of 900m2 (30m × 30m) with a residential land cover classification is estimated to attract (in) 1.0 × 0.25 × 900/100 = 2.25 trips during the morning peak, while the area will generate (out) 1.0 × 0.75 × 900/100 = 6.75 trips during the morning peak.

The resulting trip generators and attractors for the City of Cape Town in the morning peak are shown in Figures 2a and 2b, respectively. The corresponding figures for eThekwini and Gauteng are shown in Appendix C.

The trip generators reflect the population density, but it is not as pronounced as the population itself. Since many primary activities are education-related, and walking is a dominant mode to education, a sizeable portion of the population does not incur (motorised) trips.

What is notable, particularly in Cape Town and eThekwini, arguably because the areas are smaller, is that the Central Business Districts (CBDs) become high areas of travel activity: both trips attracted and generated. Yet, the population density itself does not justify the number of trips, at least not those generated.

Something else seems to be at play, and it is one of the main arguments of this study that much of our cities' transport infrastructure is based on a hub-and-spoke network with the focal point, the hub, being the economic centers.

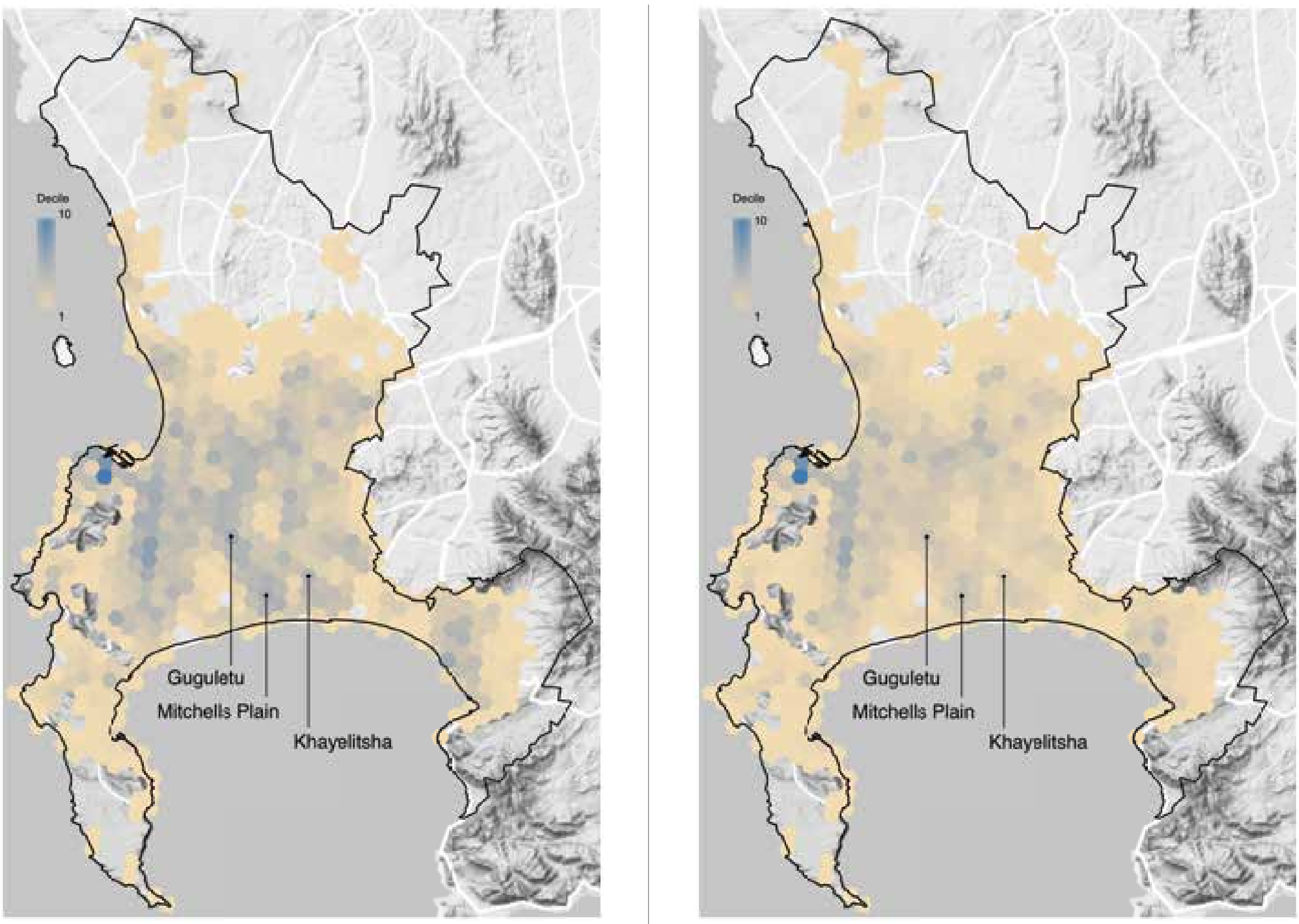
One may argue that that should indeed be the case. Possibly. But many of the peripheral poor simply use the hub to connect to job opportunities located on other spokes. This is because a) there are no direct connections between the spokes, and b) their skill set or education disqualifies them from competing for the limited jobs close to the hubs. Add to this the fact that different modes of public transport are rarely, if ever, truly integrated. More specifically, the co-location of stops does not make two modes *integrated*.

Subsequently, people turn more to private car as a viable transport alternative, albeit more expensive, to get them more directly from their origins to destinations. The captive public transport users, however, remain paying high fees for inefficient connections to get them to their employment. And the result? The increasing economic inequality we observe.

Trip estimates is one perspective to gain an understanding of how people travel, and the resulting congestion caused by motorised transport. Unfortunately both the National Household Travel Survey (NHTS) and the National Census lacks detailed locations for place of employment. Furthermore, South Africa does not have a central data set on employers, and where the companies do report their locations, it is limited to the official legal address and, at best, the head office. For multi-facility employers we therefore have a very limited idea of where people work.

[1] *More technical details of how the classification was done can be found the dataset's Data User Report and MetaData.*

Figure 2: Trip distributions for the City of Cape Town in the morning peak: a) outbound trips (generators); b) inbound trips (attractors).



3. Estimating pace of employment

If existing data sources are lacking in terms of indicating place of employment, one would naturally ask next: "so what is the alternative?" This study aimed to investigate an alternative approach to present economic activity, especially related to transport demand. More specifically, it is argued that places of employment are frequently visited by commercial vehicles. These visits include product collections, deliveries, or some service activity of a commercial nature. And with a large data set of commercial vehicle movement available in South Africa (Joubert and Axhausen, 2011) we can use the commercial activity density as a proxy for place of employment.

3.1 Extracting vehicle movement

Geospatial Positioning System (GPS) movement data is becoming ubiquitous. For this study we had access to the GPS trace data for approximately 15 000 commercial vehicles over a full year: April 2013 to March 2014. From the movement traces we extracted activities for each vehicle in the same way as Joubert and Axhausen (2011).

Methodology For each vehicle, we considered the ignition signals to indicate the start and end of an activity. That is, when the ignition is switched off it indicates the start of an activity, and when it is switched back on it indicates the end of the activity and the start of the trip towards the next activity. False starts and stops are filtered, and one ends up with a long string of subsequent activities for each vehicle. For each activity we know the detailed location, and well as the start and end times, and therefore the duration. We use the threshold duration of 300 minutes (5 hours) suggested by Joubert and Axhausen (2011) to distinguish between depot-like activities that signal the start and end of an activity chain, and all the other activities that make up the activity chain. Studying all vehicles' activities together, and applying the density-based clustering popularised by Joubert and Meintjes (2015a,b) allows us to identify the actual facilities where commercial vehicles perform their activities. If many vehicles performed activities at a particular spot over time, such locations is more likely to be identified as belonging to the same establishment. This is particularly useful as it is a plausible argument that commercial vehicles conduct activities that have economic value, be it in terms of freight collected/ delivered, or conducting a commercial support activity.

Commercial vehicle movement is the manifestation of how companies do business with one another.

3.2 Commercial connectivity

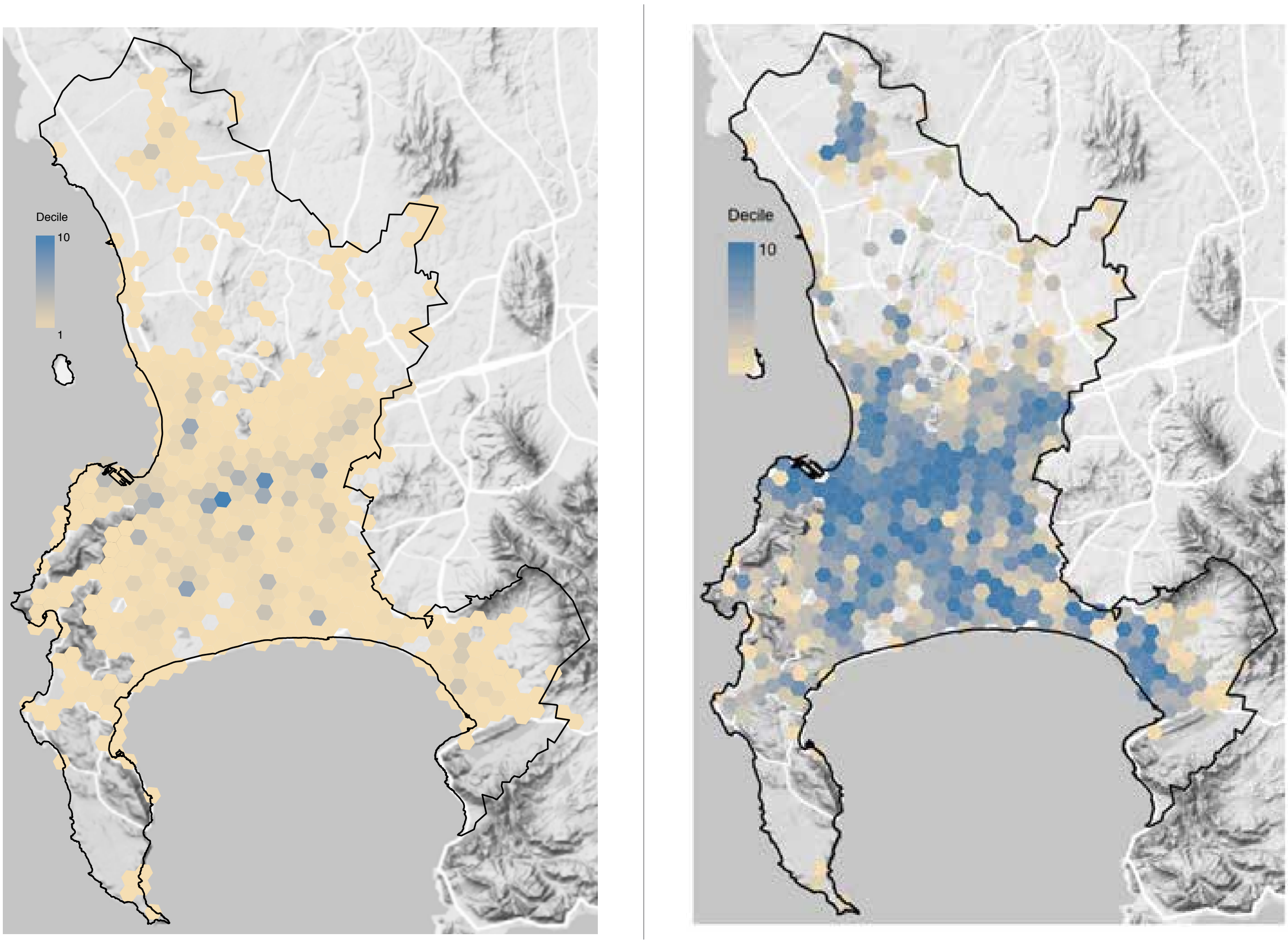
And it was this business, commercial, and ultimately economic connectivity that inspired Joubert and Axhausen (2013) to study the inter-firm connectivity using social network theory. Although we will build on their foundation, this study will make a number of valuable and novel contributions.

We use every direct trip between two facilities, say *a* and *b* observed in the activity chains, as a proxy of an economic relation between the two facilities, and specifically in the direction of the trip. That is, we argue that if a truck travel directly from *a* to *b*, there must be some relationship between *a* and *b*, albeit indirectly. But while travelling from *a* to *b* implies a relationship in the direction a→b, the reverse *a*←*b* is not automatically assumed. Recurring trips from *a* to *b*, even by different vehicles, will strengthen and reaffirm the relationship.

Instead of connecting the two specific facilities, we connect the two zones in which the two respective activities occur, and the same hexagonal grid was used to achieve comparative results. The zonal aggregation achieves two valuable goals. Firstly, it partially reduces the selection bias as it evens out the contribution of each facility over a larger area. Secondly, it eliminates privacy concerns that were introduced once unique facilities were identifiable as a result of the density-based clustering. Aggregating the commercial vehicle activities to zonal level, we observe the densities for Cape Town as shown in Figure 3a. The corresponding figures for eThekwini and Gauteng are shown in Appendix D. Each trip has a start and ending activity, and connects the two facilities with one another. We therefore deal with a network of connectivity that essentially captures the level of commercial vehicle movement—a proxy for economic activity— within a zone and, more importantly, between zones. Each zone is a node in the network, and the weighted links connecting nodes represent the level and direction of activity.

This section deals with connectedness as observed from multiple disaggregate activity chains that are not limited in its temporal scope. That is, all trips over the entire duration of a day are considered. We acknowledge that such disaggregate data may not always be available. That said, the connectivity may be calculated in a similar way using available origin-destination (OD) matrices that are often more regularly available as inputs to cities' transport models. It goes without saying that we specifically refer to the freight/commercial-related OD matrices.

Figure 3: Economic hotspots as a function of commercial vehicle activities in the City of Cape Town: a) commercial vehicle activities; b) closeness centrality.



3.3 Centrality and closeness

In network theory the notion of *centrality* is used to calculate the relative importance of a node in the network. Some metrics, like *degree centrality* simply calculates the number of connections that a node has with other nodes in the network. Other metrics like *betweenness centrality* deals with how connected a zone is to other, well-connected zones in the city.

Of particular interest in this study is the metric of *closeness*. The closeness of a zone is calculated as the sum of the length of the shortest paths between that zone and all other zones in the graph. In its normalised form it is actually the *average* length of the shortest paths. So, for zone i the (normalised) closeness is calculated as

$$C(i) = \frac{N}{\sum_j d(i,j)}$$

where N is the number of zones in the graph, and $d(i,j)$ is the (graph) distance between nodes i and j. But we use a weighted network where a link's weight is the strength with which two zones are connected. The stronger the link, the closer the two zones are to one another. We use the generalisation of Opsahl et al. (2010) to calculate the (weighted) shortest distance between two zones i and j, denoted by $d^{\omega\alpha}(i,j)$, as

$$d^{\omega\alpha}(i,j) = \min\left(\frac{1}{(w_{ih})^{\alpha}} + \cdots + \frac{1}{(w_{hj})^{\alpha}}\right)$$

where α is a tuning parameter, and the node index h indicates all intermediary nodes between i and j.

The practical interpretation of closeness is useful. If zone *A* is directly connected to zone *B*, then it means that there are already trucks or LDVs conducting direct trips from *A* to *B*. Should an entrepreneur or small business owner wish to tap into the distribution channel through load consolidation (co-loading), for example, such opportunities exist. When considering all the zones—the sum in the denominator—one gets a sense of how *close* a zone is to all other zones. A shortest path longer than one means that goods have to be transhipped at intermediate zone(s), which may not be practical unless the business sending the goods also have some control over, or are actively involved in the operations at the intermediaries. And such involvement is unlikely for entrepreneurial and startup organisations. Alternatively, the transshipment must be done at an integrated (re)distribution centre.

A large closeness value suggests that a zone can connect easily/directly with many other zones, implying access to more, local markets. And this is where the value lies. In his proposed World Bank lecture, Dr Somik Lall argues that African cities are characterised by disconnectedness. This holds for national, regional, and local levels.

We specifically use topological (graph) distance instead of geometric (or geographical) distance. The reason being that when I have access to a vehicle to move my goods from one zone to another, I have *access* to a market. The physical distance may be of lesser importance. We acknowledge that there is a valid argument that transport cost might be more accurately accounted for using geometric distance. But transshipment also carries a not-so-easily-quantifiable cost, and such (hidden) costs are better addressed and accounted for using topological distance.

The closeness for each zone is calculated using (1) and visualised in Figure 3b. The corresponding figures for eThekwini and Gauteng are shown in Appendix D. In all three study areas we see that high closeness is scattered more evenly throughout the areas.

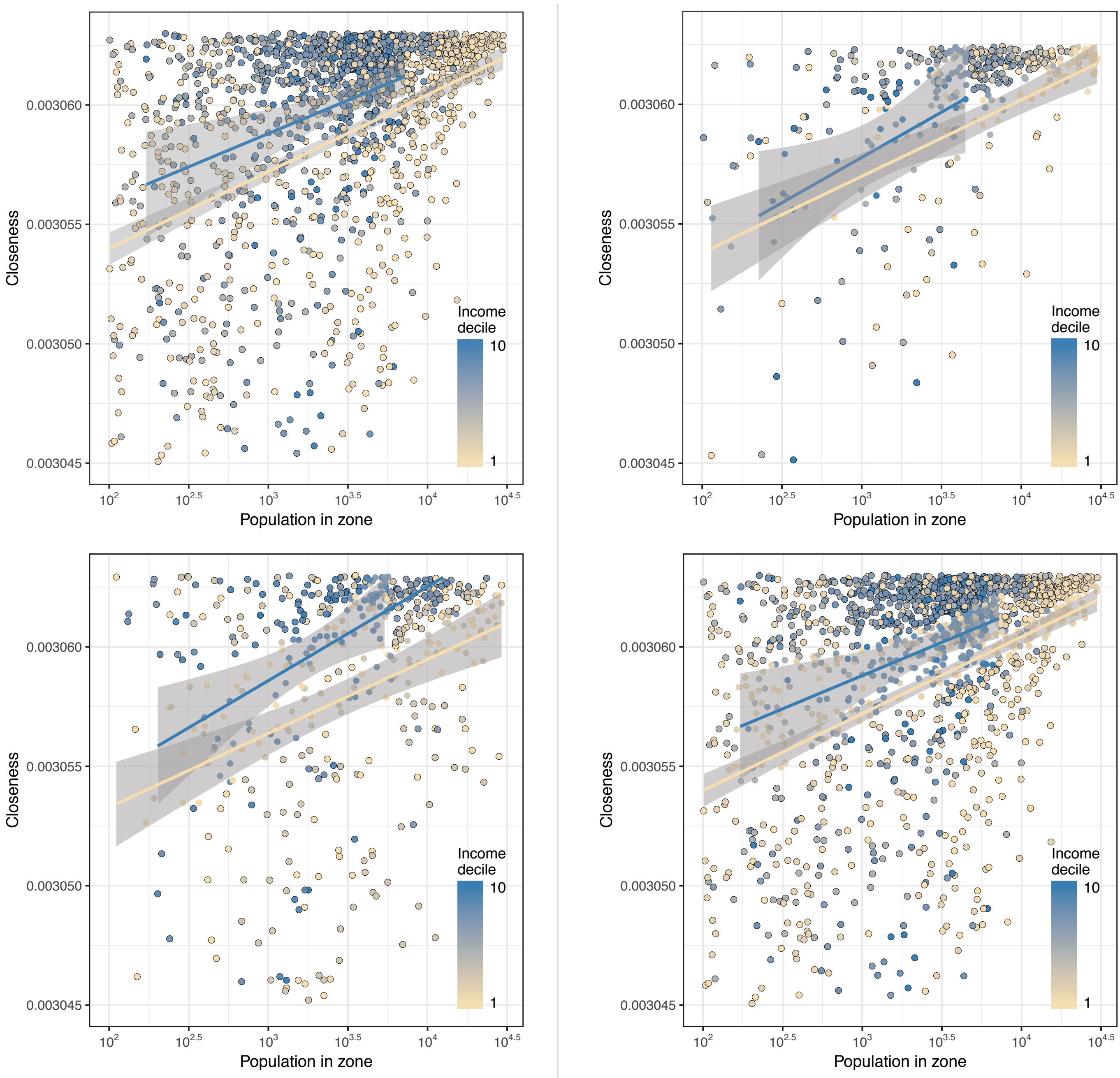
3.4 A case for alternative connectivity

One of the main contributions of this work is to show how the three different study areas perform in terms of closeness. More specifically, we show that Dr Lall's claim/conjecture indeed holds by quantifying that, generally, low-income zones are less connected, implying that they have less access to markets. This has implications for promoting entrepreneurial and township economies to develop. Why is this important? For two reasons. Firstly, if entrepreneurial township economies become sustainable, people need not travel long distances to work. Instead, the work opportunities move to where the people are. Secondly, in a country plagued with unemployment, we need solutions alternative to government being a major employer.

In the absence of better connectivity, entrepreneurial interventions may be short-lived, or worse, only remain while subsidies and fiscal support are present. To achieve lasting and sustainable economic growth, small enterprises will have to secure low-cost and efficient access to its markets. Alternatively, economic development interventions must go beyond the startup phase and also develop the (possibly shared) supply chain channels. A proposal to study the informal and township logistics networks has been presented by the author.

In Figure 4 the closeness of the three study areas are compared. Each dot represents a zone, and its position shows its closeness (y-axis) as a function of the population in that zone (x-axis). The colour of each dot depicts the median income decile for the particular zone.

Figure 4: Explaining closeness as a function of household size and income. Trends for the upper and lower quantiles are shown with its respective 95% confidence intervals: a) Cape Town, b) eThekwini and c) Gauteng.



In each plot we also fit a linear model to the upper income quintile (two deciles, thus 20%) in blue, and the lower income quintile in the wheat colour. A couple of observations can be made from the figures. Firstly, the overall closeness of the City of Cape Town is lower than the other study areas. This can be attributed to the geography where the ocean, limiting connectedness for road-based transport, bound a large portion of the city.

Secondly, high-income zones are better connected than low-income zones. This is seen in the linear function for the rich (blue line) being above that of the low-income zones (wheat colour line). This is true for all three study areas. Quite clearly in eThekwini, and especially Gauteng, there are clusters of low-income zones (wheat colour dots) with lower closeness values that appear below the respective clusters of high-income (blue) zones.

Whether deliberate planning regimes cause the differences is not clear-cut. It is a plausible argument that the interaction of housing and transport planning in planning authorities are too isolated. In Cape Town's benefit is its own *MyCiTi* rapid transit, with only one other commercial bus transit provider, *Golden Arrow*. The network design is arguably more driven by connectivity than myopic, single trip, subsidy-dependent commercial gain. A deliberate attempt to trade off the conflicting housing and transport objectives seems to pay dividends in Cape Town. It is argued that from an equity perspective low-income households should be closer to economic opportunities as they can less afford the transport to these opportunities.

A second contribution of this project is quantifying closeness and showing (Figures 3b, 15b and 16b) that well-connected zones are more evenly spread throughout the study areas. We believe this supports a call for transport authorities to not focus so much on trunk and corridor transit lines. Such efforts result in hub-and-spoke networks with limited connectivity.

Although these trunk routes are necessary, they should be supported with a trunk-connecting feeder network. This will allow individuals to search more widely, and access more job opportunities.

But such an integrated network needs to be designed. It cannot be expected that other service providers, most notably the paratransit minibus taxis, will chip in and fill in the gaps in a way that benefits the commuter. Research has shown that the association-based paratransit with its commercial gain objective evolves into having a network design that serves their commercial purpose more than what it serves the connectivity of the commuters (Neumann, 2014).

4. Travel behavior

We now return to looking at how people behave in response to the available transit options. To do that we unpack the revealed travel behaviour as reported in the NHTS.

Transport authorities are continuously faced with challenges on where to invest in infrastructure. The intent is very noble: to provide citizens with mobility so they can participate in and contribute to the economy. But spending the fiscus on infrastructure is not simple at all, and governments are always burdened with difficult trade-offs. On the one side they are faced with the question *"who gets the benefit of the infrastructure?"* Indeed, this is not easily identifiable, yet on the other side they are faced with as tough a question, if not more controversial: *"who **pays** for those benefits?"*

Government's dilemma has been highlighted in recent years, in South Africa specifically. Very large amounts were spent on transport-related projects for which the returns are challengeable. The Gautrain, a rapid rail link connecting Tshwane (Pretoria), Johannesburg and the O.R. Tambo International Airport, has proven one of the more successful stories even though the service has not nearly reached the promised (predicted) ridership. Government has to foot the bill for low ridership in the form of *Patronage Guarantee* that is payable to the Concessionaire so they can achieve their *Minimum Required Total Revenue*.

Bus Rapid Transit (BRT) systems have been built in a number of metropolitan areas, costing billions of South African Rand. Some of these, like in Nelson Mandela Bay, is inoperable, while the likes of Tshwane's *A Re Yeng* runs mainly empty. The City of Cape Town's *MyCiTi* is one of the more successful ones in terms of service coverage, frequency and connectivity. Many of these services seem to remain stuck in their early stage development. Although the systems are sold to the public, and funders, based on its well-connected networks in the final implementation phase, the projects rarely move beyond the first few phases. This is often due to low ridership along corridors, and the resulting (lack of) fare revenue, not justifying further development.

An example of a dedicated study to understand commuter choice behaviour is that of Venter (2016). The study sets out to explain, at least partially, why the current, infrastructure heavy approach to BRT implementations in South Africa yield lower passenger demand, poorer financial performance, and higher subsidy reliance. This was achieved by studying the preferences and choice behaviour, combining stated and revealed preference surveys.

Transit-oriented development and the argument of *"if we provide the infrastructure, then people will use it and develop around it"* does not seem to hold that well.

It is for this reason that researchers and practitioners spend a lot of effort on understanding people's motivation for choosing specific modes. The paratransit mode, commonly referred to in South Africa as *minibus taxis*, is often made out to be very unsafe, expensive, and erratic. Yet, despite the efforts to formalise the mode, it keeps growing without any subsidy. People choose to use the mode. One can argue that they're captive and don't have much of a choice. Be it as it may, the mode grows.

4.1 Data preperation

To try and understand how and why people make specific travel choices, data is gathered through surveys. The sole data set that is the focus of this study is the NHTS. Its aim was to gain insight into the travel patterns and transport problems faced by South Africans. The specific objectives were to a) serve as the basis for South African Department of Transport (DoT) research, planning and policy formulation; b) assist transport authorities to affectively target subsidies; and c) serve as a data source for the definition and measurements of Key Performance Indicators (KPIs).

The questionnaire type survey was completed by 155,041 individuals that were associated with 42,221 households. This section describes the data cleaning and pre-processing phase we conducted to get the survey data into a more usable form. For purposes of visibility, repeatability and reproducibility, we used R for data preparation and cleaning (R Core Team, 2017). The complete procedure is captured in the markdown document ChoiceDataPrep.Rmd, which is available on request.

We split the entire data set into those individuals who indicated they travelled to work as their primary activity, referred to as WORKERS (32,345 observations), and those who indicated they travelled to education (any level), referred to as SCHOLARS (42,448 observations).

The explanatory variables we want to consider as covariates, also referred to as *predictors*, that is, variables that can possibly help predict an individual's main mode of transport, are categorised into individual, household, and trip or travel-related variables.

4.1.1 Individual characteristics

The individual characteristics we were interested in include age, gender, race, level of education completed and whether the individual holds a license.

Although the survey asked the actual age in years completed, we aggregated it to age groups and reports the frequencies for the two data subsets in Table 1. For each variable, the table caption indicates the encoding used for that variable, and the table reports on the encoding used for each discrete category within that variable. Observed gender frequencies are reported in Table 2.

Table 1: Age group (ageGroup).

		Percentage of observations	
Description	**Encoding**	**Scholars**	**Working**
Infant (<6 years)	infant	16.9%	-
Child (6–12 years)	child	40.7%	-
Young (13–23 years)	young	39.8%	7.7%
Early career (24–45 years)	early career	2.2%	62.9%
Late career (46–65 years)	late career	0.2%	27.4%
Retired (>65 years)	retired	0.2%	1.9%

Table 2: Gender (gender).

		Percentage of observations	
Description	**Encoding**	**Scholars**	**Working**
Female	female	46.5%	44.0%
Male	male	50.5%	56.0%

Next we report the observed race distribution in Table 3. Table 4 reports on the level of education completed by the individual. The 29 classes of education used in the NHTS, Question 3.1, were aggregate to those five reported in the table. When we indicate primary we imply primary school, which is grades R through 7, *or any part thereof*.

Finally we look at whether an individual has a license. The question was asked to all people 16 years or older and distinguished between motorcycle, light vehicle (car) or heavy vehicle. Since we do not include motorcycle as a specific mode (too few observations), we ignored the corresponding license type and aggregated the remaining two. In South Africa, if you have a heavy vehicle license, you are allowed to drive light vehicles too, qualifying you therefore for private car as a mode. The results are reported in Table 5.

Table 3: Race (race).

		Percentage of observations	
Description	**Encoding**	**Scholars**	**Working**
African/Black	african/black	87.1%	69.3%
Coloured	coloured	80.0%	14.1%
Indian/Asian	indian/asian	1.3%	3.8%
White	white	3.6%	12.1%

Table 4: Completed education (edu).

		Percentage of observations	
Description	Encoding	Scholars	Working
No education completed	none	20.8%	5.1%
Primary school	primary	48.8%	15.0%
Secondary school	secondary	29.7%	61.7%
Tertiary	tertiary	0.7%	15.9%
Postgraduate	postgraduate	0.1%	2.4%
Retired (>65 years)	retired	0.2%	1.9%

Table 5: License ownership (lic).

		Percentage of observations	
Description	Encoding	Scholars	Working
Have license	true	1.3%	39.5%
No license	false	98.7%	60.5%

4.1.2 Household characteristics

Although some questions were asked about the households as a whole, we associate those characteristics with each individual and report them here at individual level. The first variable is location. Although many households did not report their activity location (work or education) at the Transport Analysis Zone (TAZ) level, the provincial code was usually captured. In Table 6 we report the location mainly at provincial level, but include two specific metropolitan areas, City of Cape Town (Western Cape) and eThekwini (KwaZulu-Natal), as they were specific focus areas in this study. For Gauteng, along with the two metros, we used detailed TAZ codes to identify them more specifically. We did not use location as a covariate. Instead we used it to split the data set later and develop location-specific modal choice models.

Table 6: Location (province).

		Percentage of observations	
Description	Encoding	Scholars	Working
Western Cape	wc	7.8%	14.7%
City of Cape Town	ct	4.0%	7.4%
Eastern Cape	ec	17.8%	10.3%
Northern Cape	nc	1.3%	4.7%
Free State	fs	7.0%	6.5%
KwaZulu-Natal	kzn	24.6%	18.6%
eThekwini	et	4.6%	7.5%
NorthWest	nw	7.2%	6.2%
Gauteng	gt	13.3%	25.5%
Mpumalanga	mp	8.3%	7.4%
Limpopo	l	12.8%	6.1%

The second household characteristic we were interested in is household size. Although not explicitly asked, we could infer it since each individual has a unique household number. We aggregated the number of household members as per Table 7. The aggregation can probably see a 'medium' class added as well, but there was no specific evidence to support it. The data does unfortunately not indicate the role, so a newly married couple with no kids and a single mother with one child will both be (erroneously) classified as a couple. Further refinement can be made to infer roles from the age group in future work.

Table 7: Household size (hhSize).

		Percentage of observations	
Description	**Encoding**	**Scholars**	**Working**
Single	single	0.7%	11.3%
Couple (2)	couple	3.9%	17.1%
Small (3-5)	small	47.8%	49.2%
Large (>5)	large	47.6%	22.4%

Household income was considered next. The NHTS imputed income from various sources and although an actual South African Rand value was assigned to each household, we choose to aggregate it into the same classes as used by the University of South Africa's Bureau of Market Research. The results are shown in Table 8.

Table 8: Household income (hhIncome).

			Percentage of observations	
Description	**Monthly ceiling (ZAR)**	**Encoding**	**Scholars**	**Working**
Poor	4 530	poor	7.8%	14.7%
Low middle class	12 644	low middle class	17.8%	10.3%
Emerging middle class	30 328	emerging middle class	1.3%	4.7%
Middle class	52 394	middle class	7.2%	6.2%
Upper middle class	71 992	upper middle class	13.3%	25.5%
Emerging affluent	110 820	emerging affluent	8.3%	7.4%
Affluent	>110 820	affluent	12.8%	6.1%

Arguably related to income is the dwelling type of the household. We include this as it was explicitly asked in the survey, and because it may be a good proxy for the household's asset value. The aggregated results are shown in Table 9.

Table 9: Household dwelling type (hhDwell).

		Percentage of observations	
Description	**Encoding**	**Scholars**	**Working**
Formal (semi) detached house	formal	17.5%	72.6%
Cluster housing in complex	complex	0.5%	1.9%
Appartment in block of flats	block	1.8%	4.5%
Room/flat/dwelling/backyard	backyard_formal	1.7%	3.6%
Informal dwelling backyard	backyard_informal	2.2%	4.0%
Traditional	traditional	16.8%	5.3%
Informal	informal	5.3%	8.0%
Large (>5)	large	0.1%	0.2%

The final household covariate is access to vehicles. We distinguish between cars and bicycles. Although the survey asked for the number of vehicles the household has access to for private use, as a driver, we ignore the joint decision-making among household members and simply indicate whether the specific vehicle type is available to the household or not. The results are shown in Table 10.

Table 10: Household access to vehicles (hasBike for bicycles and access for car).

			Percentage of observations	
Vehicle type	**Description**	**Encoding**	**Scholars**	**Working**
Bicycle	has access	true	2.0%	2.7%
	no access	false	98.0%	97.3%
Car	has access	true	25.8%	43.2%
	no access	false	74.2%	56.8%

4.1.3 Travel characteristics

Travel choices are not only made based on one's socio-demographic profile, but on the trip's travel characteristics as well. Essentially the NHTS captures revealed preferences in that much of the travel characteristics are a *result* of the chosen mode, and are not necessarily *causing* the specific mode to be chosen. Be it as it may, we have to start somewhere and any systematically collected data is better than nothing.

One of the questions in the survey asked travellers when they left home and started their journey. We used this to determine whether the journey started before, during or after the peak period. The results are shown in Table 11. A similar question was asked about their arrival time at either work or place of education. Adding the questions that dealt with walking time to and waiting time for the first mode of transport used, as well as the final walking leg, and we can estimate both walking time and in-vehicle time. The sum of these gave us the total travel time. The inferred total travel times were aggregated into deciles (based on the WORKERS data).

Table 11: Journey start time (peak).

		Percentage of observations	
Description	**Encoding**	**Scholars**	**Working**
Before 06:00	before	2.5%	20.1%
06:00 - 09:00	peak	95.7%	74.4%
After 09:00	after	1.8%	5.1%

Using estimated peak period speeds of 2.5km/h for walking; 10km/h for cycling; 20km/h for bus[2] ; 5km/h for train; and 40km/h for taxi and car, total travel distance could be estimated since we already distinguish in the travel time between walking and in-vehicle travel time. Like travel time, the distance was aggregated into deciles. The resulting time and distance distributions are shown in Table 12.

[2] *This is higher than international estimates of below 20km/h, but the majority of bus services provided are not schedule-based, but rather point-to-point services during the peak period.*

Table 12: Journey time and estimated distance (timeF and distF respectively).

		Percentage of observations	
Description	**Encoding**	**Scholars**	**Working**
Time < 14min	1	15.3%	10.0%
14-20min	2	27.1%	15.2%
20-30min	3	27.2%	23.8%
30-35min	5	2.2%	2.0%
35-45min	6	11.0%	10.5%
45-60min	7	11.6%	18.5%
60-90min	9	5.5%	11.2%
>90min	10	-	8.8%
Distance <0.8km	1	23.2%	10.0%
0.8-1.3km	2	28.9%	11.2%
1.3-3.8km	3	20.8%	10.4%
3.8-8.5km	4	8.0%	14.6%
8.5-10.5km	5	5.4%	10.9%
10.5-16.7km	6	5.7%	15.5%
16.7-25.0km	7	1.4%	3.8%
25.0-29.2km	8	3.80%	8.5%
29.2-39.5km	9	2.3%	8.8%
>39.5km (max 254km)	10	1.4%	6.1%

For educational trips the survey asked about three possible consecutive modes used, while for work trips four possible modes were considered. From this we could determine the number of transfers. But mode combinations had to be simplified into a single main mode[3]. To do this a simply heuristic was applied. We denote with train ≻ bus that if both train and bus were used in the journey, the bus was probably used to get to the train, which was subsequently used as the main mode. Similarly, bus ≻ taxi implies that when taxi and bus is used, the taxi was likely used as used as feeder mode.

In general then we specify the heuristic as

train ≻ bus ≻ taxi ≻ passenger ≻ car≻ bicycle ≽ walk

noting that bicycle ≽ walk implies the two modes may have equal weight, but one is still likely to first walk to the point where you parked/locked your bicycle. One exception was, given the flat, cash-based fare structure of the minibus taxis, that if multiple taxi modes were chosen along with bus or train, that taxi would be identified as the main mode.

Finally, we report the observed modal choices in Table 13.

Table 13: Observed main mode (mainMode).

		Percentage of observations	
Description	**Encoding**	**Scholars**	**Working**
Walking	walk	71.8%	28.4%
Cycling	bicycle	0.1%	0.8%
Bus	bus	4.7%	6.4%
Train	train	0.7%	3.0%
Taxi	taxi	11.8%	23.1%
Car, as passenger	passenger	10.2%	11.1%
Car, as driver	car	0.7%	27.3%

[3] *The proposed modelling approach allows for future work in which multiple modes and trip-chaining can be studied.*

Traditional discrete choice models are based on a utility-maximisation framework. Within this framework we assume that every commuter, the decision-maker, has perfect knowledge of all the available alternatives.

The NHTS, however, does not ask about the alternative modes, let alone attributes associated with it. So if one wants to estimate a discrete choice model, you need to infer mode-specific characteristics for each of the non-chosen alternatives. This in itself introduces a level of bias into the data set that one wishes to use to estimate a model from. Using a model to impute data for yet another model can render the results very unusable very quickly.

Now add the complexity of (accurately) inferring what finite set of alternatives are really available for each decision-maker, and we have ourselves a very messy choice model very quickly.

4.2 Choice modelling

We want to understand *causal structure*. That is, what causes people to make specific decisions? The classical, discrete choice models assume that the decision-maker has the rational capability, capacity, and time to evaluate all the alternatives in its entirety before making a decision. The evaluation is based on utility-maximisation, and the decision-maker will consistently pick the alternative yielding the highest utility. Discrete choice modelling assumes that the decision-maker will make the same choice every time under the same conditions. We refer to that as deterministic decision-making[4].

As modellers it is our goal to try and specify our models, and estimate it's parameters so that we can predict those modal choices. Since the data we use to estimate the choice model is imperfect and incomplete, the unobserved effects are accounted for through a random error term. The fact that people, in reality, may make different mode choices under the same conditions is referred to as *modal choice heterogeneity*. Ma et al. (2017) conclude that such complexities in human decision-making have led to decision rules being used in recent state-of-the-art research as this may yield more practical solutions. They, with other recent studies like Wu and Yang (2013) and Ma (2015), found that a Bayesian network (BN) not only performed on par, or even outperformed its multinomial logit counterparts, but revealed (intuitive) insights into the choice process that traditional models were incapable of.

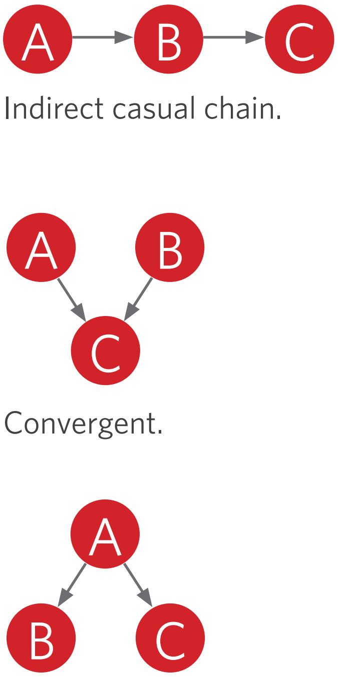
4.2.1 Bayesian network

A BN, also referred to as a *belief network*, is an approach to model causal relationships between a set of variables in some decision-making domain. These causal relationships between variables are presented by conditional probabilities. This means that we can indicate a kind of *if-then* rule under uncertainty (Ma, 2015).

The relationships are established through inductive knowledge discovery and combines Bayesian probability theory and graph theory. So, more formally, we can describe a BN as an acyclic directed graph S($\boldsymbol{X},\boldsymbol{A}$), where $\boldsymbol{X}$ is a set of nodes and $\boldsymbol{A}$ is a set of arcs. Each node $X_i \in \boldsymbol{X}$ represents one of the explanatory variables (covariates), with one node being the response variable, i.e. main mode chosen. We use capital X_i to indicate that the covariates are essentially random variables. Each directed arc between a pair of variables denotes a direct causal relationship, in the direction of the arc, between the two variables it connects. For example, say an arc connects variables *A* with *B* where A,B $\in \boldsymbol{X}$, we denote the arc as *A* → *B*, and it implies that A directly influences *B*. The variable *A* is said to be the parent of *B*, and likewise *B* is said to be the child of *A*. The set of parents for node $X_i \in \boldsymbol{X}$ is denoted with $pa(X_i)$.

Three basic types of connections are observed in the BN structure and are illustrated in Figure 5. The indirect causal chain in Figure 5a implies that *A* has an indirect effect on *C* via variable *B*. The convergent relationship shown in Figure 5b indicates that *C* is a common effect of both *A* and *B*. Finally, the divergent relationship shown in Figure 5c indicates that *A* is a common cause for *B* and *C*.

Figure 5: Basic relationships among variables (adapted from Ma et al. (2017)).



Indirect casual chain.

Convergent.

Divergent.

[4] *Look at studies from psychology and cognitive theory that discusses the limits of rational choice. Ma et al. (2017) cites Newell and Simon (1972) and Rubinstein (1998).*

If we order the nodes $X_1, X_2,...,X_n$, which is implied by a BN being an *acyclic* graph, we can express the joint probability distribution compactly using the chain rule in (3).

$$P(X_1,X_2,...,X_n) = P(X_1\ ) \times P(X_2\ X_1) \times \ldots \times P(X_n\ |X_1,\ X_2,\ ...,\ X_{n-1})$$
$$= \prod^n{}_{i=1} P(X_1\ |pa(_1))$$

And it is this property that makes BNs intuitive (and we argue therefore valuable) to understand the complex process of modal choice. For example, using the indirect causal chain of Figure 5a and applying it to modal choice, consider license → peak → time → modalChoice. Assume having a license is a random variable. *Not* having a license increases the probability that a person will travel during peak period when public transport services are available, or more frequent. Travelling during the peak is likely to increase your travel time, which, in turn, influences your chosen mode.

One of the many benefits of using BNs is that one can easily incorporate new variables or elements with little effort. For example adding age as an explanatory variable. Also, one is able to capture and represent the potential relationships jointly. For example, what is the probability of walking to school if you are a child and of coloured race?

Another benefit of BNs is that the causal structure learned from the empirical data is exploratory. That is, it is inferred from the data instead of being pre-specified in confirmatory models.

4.3 Model specifications

To be able to infer modal choice from a BN we have to find the structure of the network that best matches the joint probabilities that we observed in the NHTS data.

4.3.1 Structure learning

Three different methods of learning the BN structure have been proposed (Ma et al., 2017). The first is based on expert knowledge alone. All cause-effect relationships among variables are defined by the expert's domain knowledge. A direct consequence is that the quality and relevance of the results is highly dependent on the expert's knowledge.

A second method is a fully automated approach. No expert inputs are used, and only a goodness-of-fit measure is applied to the given data set on which the network is trained. The challenge with this method is that an expert finally may need to choose between multiple non-unique network structures.

The third method, a hybrid of the first two, is the one that we apply in this study, and is also used by Ma et al. (2017). The model is influenced insofar past research can confirm causal relationships between certain variables. All non-specified relationships are then left for the learning algorithm to discover.

There is a variety of learning algorithms available, and is widely covered in literature. In this study we use a simple yet affective, score-based, hill-climbing search algorithm. Since the algorithm is heuristic, and therefore prone to being stuck in local optima, we perform 100 restarts and choose the best scoring instance. The scoring is done using Bayesian Information Criterion (BIC) popularised by Schwarz (1978) and is a good trade-off between the complexity of the resulting network structure, and the goodness with which the network fits the empirical NHTS data.

We considered three model structures. In each case we prohibit non-causal relationships to ageGroup, race and gender. That is, for example, none of the covariates can cause your gender to change. Similarly, you may grow a few extra gray hairs while waiting for a taxi transfer, but it does not cause you to change ageGroup. Other than that, the following specifications were imposed.

v1 In the first version we imposed expert-based constraints on the BN based on prior research. More specifically we adapted the findings of Ma et al. (2017, Figure 3), which, in turn, is influenced by the review of De Witte et al. (2013) on modal choice factors. In this version of the model certain causal relationships were enforced, although the *direction* of the relationships were left for the model to discover. For example, consider the relationship between household size, encoded as family, and household income, encoded as income, with the relationship denoted by family ↔ income. One can argue that larger households, say couple instead of single, causes (larger) household income as more members are able to contribute. Another plausible argument is that larger households, say large instead of couple imply (lower) income as member(s) of the household have to look after the children and cannot contribute anymore. In either case we end up with a directed causal relationship family → income. Conversely, one can argue that household income causes the size of the household: the more money we earn as a family, the more children we can have and afford. This will result in the directed causal relationship income → family. Exactly one of these two directed relationships must be present in the final network structure. Bidirectional relationships that were included in this version were ageGroup ↔ income, family ↔ income, access ↔ income, access ↔ family, and timeF ↔ peak. Past research also suggested directional constraints. These specified causal relationships that must appear in the final network. In our case they included gender → mainMode, ageGroup → mainMode, income → mainMode, access → mainMode, hasBike → mainMode, and timeF → mainMode. The explanatory variables used only included those to which Ma et al. (2017) explicitly referred to, namely ageGroup, race, gender, lic, edu, income, hasBike, access, timeF, family, and peak.

v2 With the second version we wanted to be less restrictive than in the first. Firstly, we removed all the (bi)directional constraints that were based on past research. Secondly, we replaced the time-based covariate, timeF, with its distance-based counterpart, distF. The explanatory variables used were ageGroup, race, gender, lic, edu, income, hasBike, access, distF, family, peak, hhDwell, and transfer.

v3 The final version is only applicable to the full, national data set, as we included the location variable province. Otherwise it is exactly the same as version 2.

4.3.2 Parameter learning

Given a network structure, we estimate the conditional probability distributions using *Bayesian posterior parameter estimation* as an alternative to the classical maximum likelihood approaches. The posterior distribution of a random variable, in our case an explanatory variable, is the conditional probability that is assigned to it after all the node's parents have been taken into account.

4.3.3 Model validation

Each data (sub)set was split into a training and test set. The training set was created by randomly sampling 75% of the data, and estimating the BN structure and parameters using only the training data.

The model's ability to predict the final mode is then checked on the test set, which is the remaining 25% of the data. Here we calculate the classification error using a 10-fold cross-validation approach (Geisser, 1993). We use Bayesian predictions to predict the chosen mode using *all* explanatory variables instead of a frequentist prediction where only parent nodes are used.

4.4 Results and discussion

We start this section by reporting the overall accuracy of the models. All implementations of the BN algorithms were done using the R package bnlearn (Scutari, 2010). Table 14 reports the classification error for the three different versions of the BN structure specification. Lower values are better. We report the results for the three specific study areas as well as the national data set as a whole. A number of observations can be made.

Table 14: Classification errors for different BN structures and study areas.

	Scholars			Workers		
Study area	**v1**	**v2**	**v3**	**v1**	**v2**	**v3**
National	53.7%	54.9%	55.3%	23.1%	22.8%	23.7%
City of Cape Town	38.6%	32.2%	-	24.1%	26.8%	-
eThekwini	32.3%	38.1%	-	26.3%	26.0%	-
Gauteng	35.5%	33.6%	-	26.1%	25.0%	-

- The classification error for trips to education is consistently (much) higher at the national level than at the three metropolitan levels. One plausible argument is that at the metropolitan level the nature of educational trips are more homogenous in terms of modal options. A focused model seems to capture location-specific *mobility culture* (Joubert, 2013) better. Public transport, over and above the paratransit taxi mode, is also generally more readily available in cities. At national level, on the other hand, the modelneeds to account for both urban and rural behaviour, which is very different, resulting in higher error rates. A large proportion of education occurs in the rural areas. The concentration of work in urban areas, on the other hand, is much more pronounced, and we therefore see little difference between the national and metropolitan models. In actual fact, the national model dealing with trips to work consistently outperforms the metropolitan models, albeit slightly.
- Trips to work are predicted much more accurately, that is, with a low classification error than trips to education. This, at least initially, is somewhat counter intuitive. One could argue that trips to education is more homogenous; we have better data about the location of schools; work types vary a lot more that education types; and therefore we should be able to predict mode choice to education better. This is not the case. The best explanation is that there may be other, hidden confounding variables that influence mode choice to school that are currently not addressed in the current data set extracted from the NHTS.
- The comparison of version 1 versus version 2 network structures yields mixed results. In the majority of cases the less restrictive version 2 network structure has a lower classification error, albeit fairly small differences. For educational trips in eThekwini and nationally, and work trips nationally the version 1 network is slightly better. This could be quite insightful. Recall that the version 1 network structure is based on what past literature reported on. These causal relationships have been scientifically and rigorously scrutinised. One is tempted to answer the question *"is South Africa really that different in terms of how people commute?"* with the general notion of *"no, people are typically less unique than what they think"*. The results seem to suggest otherwise. When we remove the causal relationships that are considered

the norm internationally, we are able to predict South Africans' chosen mode more accurately, at least in some cases. As a counter argument, and since the differences between the two versions are never very big, it may be attributed simply to randomness and the sampling of the training and test sets. To verify this argument, multiple (different) samples may be drawn for the training and test sets. This will allow one to not compare single classification error values, but rather *distributions* of the errors. Estimating BNs is unfortunately computationally burdensome and such statistical experiments are left for future work.

- For the national model, adding the location variable province yields slightly worse results.

The inaccuracy of a model implies that there may be other exploratory variables that the model currently does not take into account. One reason being that such variables are not present in the data set. Very often these explanatory variables are very hard to measure and obtain in practice. For example, a person's cultural experience and sense of safety when using a specific mode could arguably prevent a young Indian teenager from travelling to school using the paratransit taxi mode, even though she lives along the street where the taxi passing her school, 3km away, drives by.

In the absence of more complete and accurate data, one can still gain insights from structural components that *do* influence people's mode choice. In the subsections that follow we look in more detail at some examples of BN structures observed.

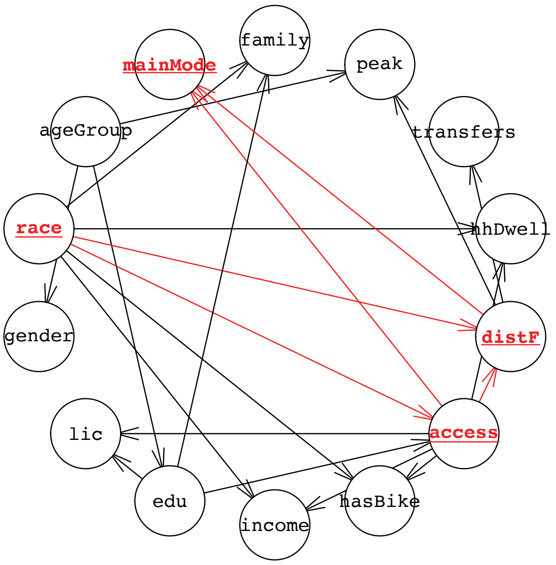
4.4.1 Interpreting causal relationships

We said that the first step in using BNs is to infer the causal relationships, the structure of the network, from the data. The learnt network for the national data for trips to education is shown in Figure 6. At first the BN looks quite complicated and highly connected. That is, there are numerous arrows, each representing a causal relationship between two variables. If one filters through the noise—as we've done by highlighting the true connectivity affecting mode choice, mainMode, in red — the network is much simpler. causal relationships, the structure of the network, from the data. The learnt network for the national data for trips to education is shown in Figure 6. At first the BN looks quite complicated and highly connected. That is, there are numerous arrows, each representing a causal relationship between two variables. If one filters through the noise—as we've done by highlighting the true connectivity affecting mode choice, mainMode, in red — the network is much simpler.

The way in which the causal relationships are interpreted is as follows. The variable we want to predict is mainMode. It is influenced directly by two variables (incoming red arrows). Firstly there is access, representing whether a family has access to a private car or not. Access to a car, in turn, is influenced by one's race.

Secondly, mode choice is influenced by distance from education, distF, which, in turn, is influenced directly by race, but also indirectly via the access variable.

Figure 6: Bayesian network of causal relationships for the national data using the v2 structure, and focussing on trips to education.



The causal relationships can now easily be *translated* into a narrative: your race influenced whether your family has access to a car or not; and having access to car directly influences your mode choice directly, but also indirectly as it allows you to search for education opportunities over larger distances.

But the causal relationships need not merely be interpreted in a qualitative narrative. Given the conditional probabilities, we can quantify the influence of each variable. For example, we can calculate what the probability is of having a car for different races. Say we want to calculate the probability of a black family having access to a car, we express it as

$$\overset{\text{The probability that}}{P}(\overbrace{\text{access} = \text{true}}^{\text{the family has access to a car}}|\overbrace{\text{race} = \text{african/black}}^{\text{given the person is african/black}})\overset{\text{is 21.1\%}}{= 21.1\%}$$

If you are coloured, the probability increases to
P (access=true | race=coloured) = 39.8%.

If you are asian or indian the probability is even higher
P (access=true | race=indian/asian) = 82.3%.

Finally, you will near certainly have access to a car if you are white, with a probability of
P (access=true | race=white) = 96.4%.

Next we show how a family's access to a car influence the children's mode choice. We denote with
P (mainMode=walk | access=false) = 80.6%
the probability that a person will choose walk as their main mode (to education, in this case) if it is given that they do not have access to a private car. When a family has access to a car, the probability drops significantly to
P (mainMode=walk | access=true) = 46.3%.

This is in line with Venter and Mohammed (2013) in their study in Nelson Mandela Bay who identified that once a family acquired a vehicle they are a lot less likely to consider/use public transport alternatives. Similarly we can calculate exactly how much the probability increases to be a passenger when a family gains access to a private car:
P (mainMode=passenger | access=false) = 4.1%
P (mainMode=passenger | access=true) = 27.2%.

The value of the BN is that one need not impose the causal structure onto the data. In- stead, the causal relationships are inferred *from* the data. Once the structure is established, the conditional probabilities are fitted. And we can study more complex relationships as well. For example, we can investigate the influence of race on choosing to walk to school if it is *given* that a family has access to a car, *and* the family resides fairly close (third decile) to the school. For blacks we denote the conditional probability as
P (mainMode=walk | race=african/black,access=true,distF=3) = 69.6%.

Whites, on the other hand, will have a very similar probability of
P (mainMode=walk | race=white,access=true, distF=3) = 69.6%.

As soon as a family has access to a car, their travel behaviour is a lot less influenced by race. But remember that there is still a strong influence that race will have on merely having access to a car in the first place.

The layered, indirect causality can be demonstrated using the causal chain race → access → mainMode. If we keep access fixed, one can determine the impact race has on modal choice. Not directly, but via an intermediate explanatory variable.
P (mainMode=bus | race=african/black,access=false) = 4.3%
P (mainMode=bus | race=white,access=false) = 9.4%

That is, for families not having access to a car, white kids are more likely to travel by bus than their african/black counterparts. Yet, when the family has access to a car, the picture changes quite a bit: african/black kids are more likely to use the bus, while there is no chance that white kids will use the bus anymore.
P (mainMode=bus | race=african/black,access=true) = 5.0%
P (mainMode=bus | race=white,access=true) = 0.0%

We can investigate what the impact would be of car access for people that have to travel further. Consider two neighbourhood friends. The one goes to the closer township school that is 2km (3rd decile) away, while the other attends the more prestigious school in town that is 20km (7th decile) away. The probability of their respective families having access to or owning a private car is calculated as follows.
P (access=true | distF=3) = 17.4%
P (access=true | distF=7) = 41.3%

And we see that the family with the child travelling further to schools is much more likely to have a vehicle available.

What is insightful is that very many of the other explanatory variables have causal relationships, influencing one another, yet having little to no (in)direct influence on the modal choice. Indeed, one can study a variety of interesting causalities, but for this study we focus on modal choice alone. The causal relationship diagrams for all the study areas, and for both trip purposes, are shown in Appendix E.

Suffice to say that BNs can be used to quantify a near-infinite number of scenarios in an intuitive way. In the next section we demonstrate how the network structure is influenced by different factors.

4.4.2 Influence of network structure

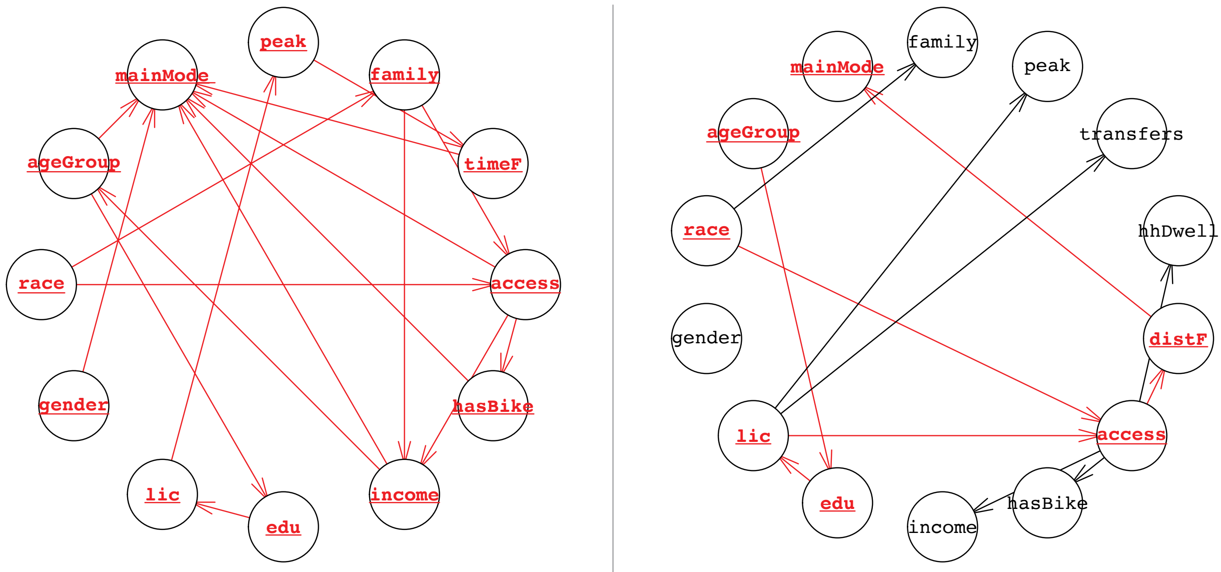
In Table 14 we report that the version 2 network structure in some cases has a lower (better) classification error.

It is worth noting the effect of the different versions. Figure 7 illustrates the different structures of the best-fit version 1 and version 2 networks. Both are based on trips to education for the City of Cape Town. In Section 4.3.1 we identified the one directional relationships that must be present in the resulting graph. All the directional constraints gender → mainMode, ageGroup → mainMode, income → mainMode, access → mainMode, hasBike → mainMode, and timeF → mainMode indeed appear in Figure 7a.

Also, as but one example, we specified the bidirectional constraint family ↔ income. Either causal relationship family → income or family ← income must be present in the final graph, and indeed, it is former.

Version 2 of the network structure specification was a lot less restrictive, and we see that being manifested in much fewer arcs in Figure 7b's network. The fewer arcs actually allows the model to reduce its classification error from 38.6% down to 32.2%.

Figure 7: Different networks resulting from different structure specifications for the same location (City of Cape Town), both focusing on trips to education.

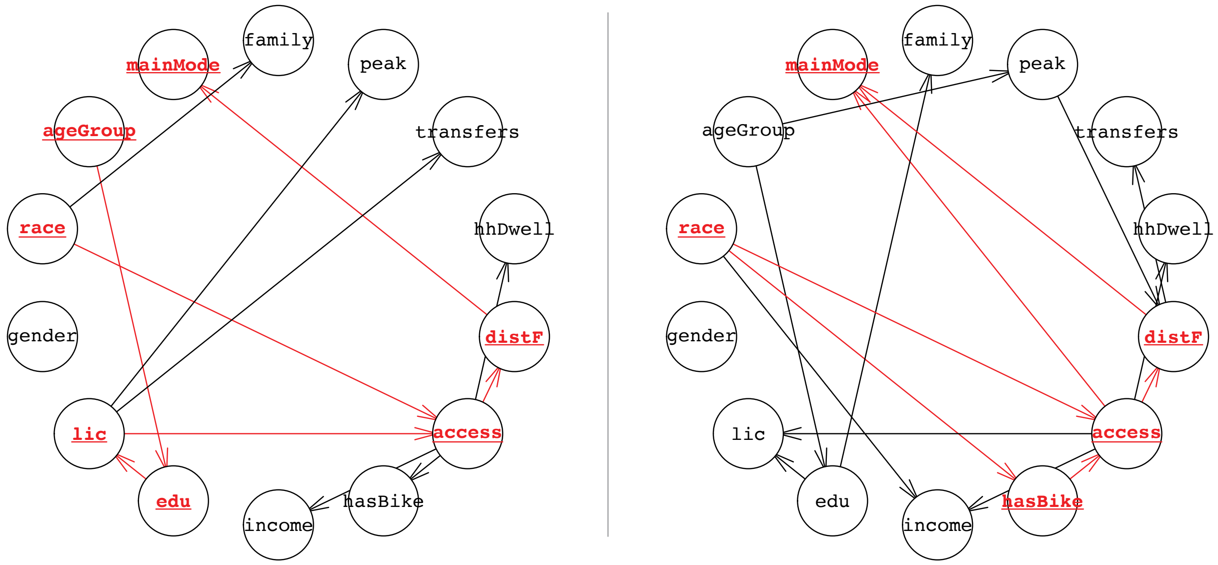


4.4.3 Influence of location

We can generate different networks for different areas by filtering on the province or studyArea explanatory variable. Figure 8 shows the resulting BNs for two such areas.

The arc access → mainMode is absent in Figure 8a, implying that, in the City of Cape Town, having access to a private car does not, at least directly, influence the mode choice. There is still an indirect influence through the causal chain access → distF → mainMode. In Figure 8b, on the other hand, the arc is present, indicating a direct influence in Gauteng.

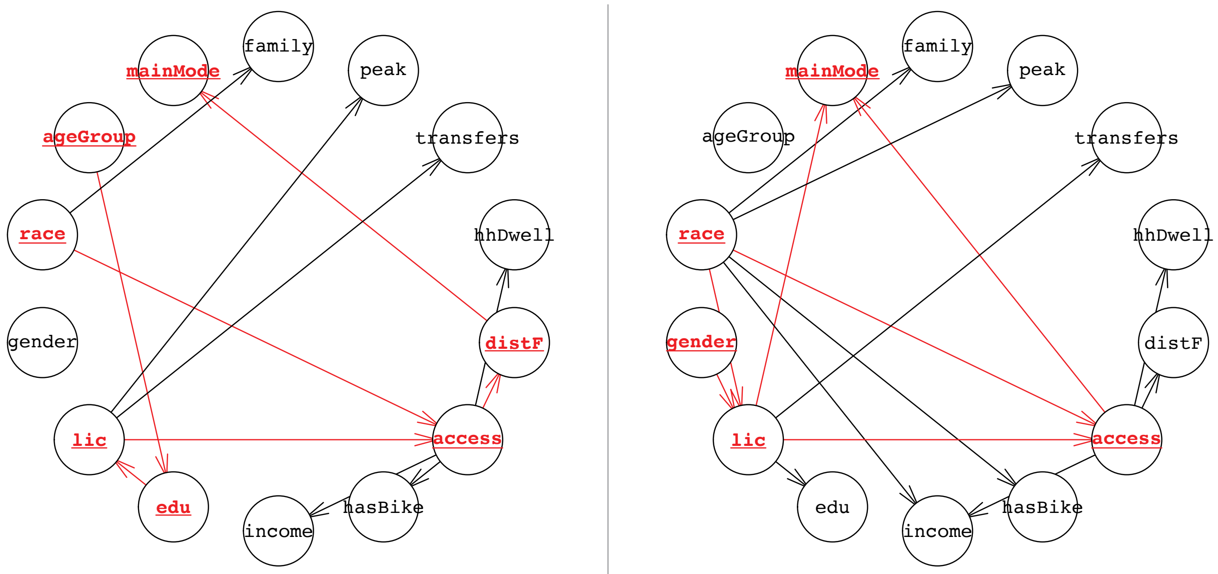
Figure 8: Different networks resulting from using the version 2 structure specification for different locations, both focusing on trips to education.



4.4.4 Influence of trip type

When we use the same network structure, but train the BN on two different data sets, in this case Scholars and Workers, we can get different networks, as is seen in Figure 9 for the City of Cape Town.

Figure 9: Different networks resulting from using the version 2 structure specification for the same location, but distinguishing between trip purposes.



4.5 Policy implications

We were able to show that the NHTS, even as a sole data source, have the potential to support very rich decision-making in quantifying the effect of causal variables. Complex causal relationships can be studied, and interpreting results can be done in a much more intuitive way than complex confounding variables and marginal utilities that are the norm in discrete choice models. With state-of-the art research in (unsupervised) machine learning, like BNs demonstrated in this study, there is room for improvement in how we advise authorities on spending their budgets on infrastructure, operations and subsidies. Better (more accurate and more intuitive) models should lead to the proverbial *"more bang for their buck"*.

4.5.1 Maintaining systematic surveys

Maintaining the systematic and periodic surveys like the NHTS is necessary, but not sufficient. This is not the standard call for decision-makers *"needing more data"*. There is a lot of modal choice heterogeneity. How people make choices about the mode to use is dependent on a lot of factors, many of which are tough to study. Even though this is true in developed countries, it is more so in South Africa with its extreme economic inequality.

To be able to capture these choice variations in an already very diverse country means that we need to sample more widely. Not deeper, asking more intricate questions, but wider. If we use survey sample sizes suggested by literature, i.e. developed countries, we may indeed find ourselves with good data, like the current NHTS, but which lacks the needed width to account for taste variation. This final call is for simplifying surveys, and rather invest in a larger footprint so that we can uncover the valuable causal relationships affecting mobility.

4.5.2 Impact of race

Somewhat surprising is the fact that neither income nor hhDwell that uses the family's dwelling type as a proxy for income, are featuring in the causal chain influencing mode choice. That is, income does not seem to be affecting one's mode choice.

One argument would be to suggest that race is unfortunately still a good proxy for income. Or more specifically, if you are poor you are more likely to be africa/black.

The causal networks differ slightly for each study area, but a recurring pattern seems to be dominant: race still plays a dominant role in one's ability to access a private car. Access to a private car, in turn, affects one's mode choice directly, but also indirectly as it allows you to search for job and educational opportunities further away from home.

More specifically, when we consider trips to work, for which the network is shown in Figure 10, we see four more explanatory variables come into play than for trips to education (Figure 6), namely ageGroup, gender, lic and peak. All variables that have an (in)direct bearing on the mode choice, as well as the causal relationships connecting them, are again highlighted in red. Having a license influences the distance people travel to work, lic → distF. Phrased differently, people having a license can consider job opportunities further away, especially since lic is a common cause (divergent relationship, Figure 5c) for both access and distF, resulting in a person then being more likely to also have access to a car for work trips.

The network structure allows us to evaluate much more indirect relationships. The effects that race, for example, have on modal choice are seen in multiple causal chains: race → access → mainMode, race → lic → mainMode, race → peak → distF → mainMode, or the even more elaborate race → lic → access → distF → mainMode. Table 15 shows the probability of mode given race and a family's access to private car, P (mainMode|race,access).

Figure 10: Bayesian network of causal relationships for the national data using the v2 structure, and focusing on trips to work.

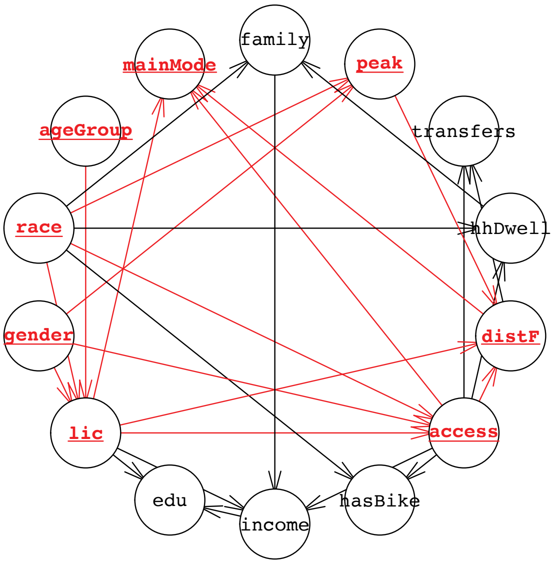


Table 15: Predicted mainMode share for different race categories, distinguishing between families without and with access to a private car.

	Car access				Chosen main mode			
Race		**walk**	**bicycle**	**bus**	**train**	**taxi**	**passenger**	**car**
African/Black	x ✓	40.6% 14.0%	1.1% 0.5%	8.5% 4.3%	3.9% 1.8%	32.3% 13.1%	13.5% 9.1%	0.0% 57.2%
Coloured	x ✓	44.8% 15.3%	1.1% 0.6%	8.0% 4.4%	3.7% 1.9%	30.3% 13.2%	12.0% 8.9%	0.0% 55.7%
Indian/Asian	x ✓	41.6% 9.9%	0.8% 0.4%	8.0% 3.0%	4.2% 1.2%	30.7% 8.0%	14.7% 7.0%	0.0% 70.6%
White	x ✓	35.4% 7.7%	1.1% 0.3%	8.8% 2.5%	4.0% 1.1%	32.8% 6.3%	17.8% 6.2%	0.0% 75.9%

We end this section linking it back to the first portion of the report that dealt with travel demand. The hub-and-spoke public transport networks provide inconvenient and limited connections to people wishing to participate in economic activities. And since the public transport is, in general, not integrated, commuters pay multiple fare structures for different service providers without there being any cap.

We sincerely appreciate that achieving fare integration (more than just payment integration allowing the use of a single payment artefact like a travel card) is no mean feat. But the alternative is much worse. The less integrated public transport is from the commuter's point of view, the more costly it is. And here we generalise cost as the monetary, time, and convenience for the commuter.

If connectivity pushes commuters to rather opt for a private car, Table 15 makes it clear what the implications are. Once a person has access to a private car, the probability of using public transport in future drops dramatically. Even though the car is of low (initial) cost and quality. And all the costly investment in establishing and maintaining public transport services will then be, effectively, wasted and a mere burden on government.

But one can only motivate people in one of two ways: the proverbial carrot or the stick. Limiting (or taxing) car ownership will, we argue, have limited effect. Especially if not supported by well-connected and frequent public transport services that are truly integrated.

The extreme alternative would be to not need to move so many people to work, but rather take employment opportunities to them, or better still, facilitate the process of people generating their own employment opportunities. And to achieve this we provided the accessibility metric, *closeness*, to guide authorities on where to focus their effort in supporting entrepreneurship.

5. Conclusion

Despite much effort to alleviate poverty in (Southern) African countries, economic inequality keeps rising. One argument is that communities are not well connected to markets.

If people cannot easily gain access and connect to the activities they wish to participate in, they may indeed look for alternatives. If their disconnectedness is because of inadequate public transport network design, they may opt for and acquire a private car. When this happens, we may lose them altogether for future public transport patronage.

Alternatively, people may wish to generate their own opportunities for work and livelihood. For this to succeed sustainably, they will need access to the markets. To this extent, this study contributes by quantifying the connectedness through a *closeness* measure. The measure is demonstrated for three study areas in South Africa. All three areas confirm the argument that low-income urban areas are less connected, and are therefore *further* from market access, than their more affluent counterparts. By no means do these analyses suggest authorities should target well-connected areas in an attempt to level the playing field. On the contrary, it should be interpreted as a strong message to elevate the less-connected areas.

References

Blaise, P. (2011). Perverse Cities: Hidden Subsidies, Wonky Policy, and Urban Sprawl. UBC Press.

Committee of Transport Officials (2013). South African Trip Data Manual. South African Na- tional Roads Agency Limited, 1.01 edition.

De Witte, A., Hollevoet, J., Dobruszkes, F., Hubert, M., and Macharis, C. (2013). Linking modal choice to motility: A comprehensive review. Transportmetrica A: Policy and Practice, 49:329–341.

Geisser, S. (1993). Predictive inference. Number 55 in Monographs on Statistics & Applied Probability. Chapman and HAll/CRC.

Joubert, J. W. (2013). Gauteng: Paratransit - perpetual pain or potent potential, volume 1 of Lecture Notes in Mobility, chapter 6, pages 107–126. Springer Berlin Heidelberg.

Joubert, J. W. (2018). Synthetic populations of South African urban areas. Mendeley Data, v1. Available online from http://dx.doi.org/10.17632/dh4gcm7ckb.1.

Joubert, J. W. and Axhausen, K. W. (2011). Inferring commercial vehicle activities in Gauteng, South Africa. Journal of Transport Geography, 19(1):115–124.

Joubert, J. W. and Axhausen, K. W. (2013). A complex network approach to understand com- mercial vehicle movement. Transportation, 40(3):729–750.

Joubert, J. W. and Meintjes, S. (2015a). Computational considerations in building inter-firm networks. Transportation, 42(5):857–878.

Joubert, J. W. and Meintjes, S. (2015b). Repeatability & reproducibility: Implications of using GPS data for freight activity chains. Transportation Research Part B: Methodological, 76:81–92.

Ma, T.-Y. (2015). Bayesian networks for multimodal mode choice behavior modelling: a case study for the cross border workers of Luxembourg. Transportation Research Procedia, 10:870– 880.

Ma, T.-Y., Chow, J. Y. J., and Xu, J. (2017). Causal structure learning for travel mode choice using structural restrictions and model averaging algorithm. Transportmetrica A: Transport Science, 13(4):299–325.

Neumann, A. (2014). A paratransit-inspired evolutionary process for public transit network design. PhD thesis, FG Verkehrssystemplanung und Verkehrstelematik, Technical University of Berlin.

Opsahl, T., Agneessens, F., and Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks, 32(3):245–251.

R Core Team (2017). R: A language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2):461–464. Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R package. Journal of Statistical Software, 35(3):1–22.

Sun, L. and Erath, A. (2015). A bayesian network approach for population synthesis. Trans- portation Research Part C, 61:49–62.

Venter, C. J. (2016). Are we giving BRT passengers what they want? User preference and market segmentation in Johannesburg. In Proceedings of the 35th Annual Southern African Transport Conference (SATC 2016), pages 658–672.

Venter, C. J. and Mohammed, S. (2013). Estimating car ownership and transport energy consumption: a disaggregate study in Nelson Mandela Bay. Journal of the South African Institution of Civil Engineering, 55(1):2–10.

Wu, J. and Yang, M. (2013). Modeling commuters' travel behavior by bayesian networks. In 13th COTA International Conference of Transportation Professionals (CICTP 2013), volume 96 of Procedia - Social and Behavioral Sciences, pages 512–521.

Apendix A - Acronyms

AADT Annual Average Daily Trip generation rate

BIC Bayesian Information Criterion

BN Bayesian network

BRT Bus Rapid Transit

CBD Central Business District

CHPC Centre for High Performance Computing

DoT South African Department of Transport

GPS Geospatial Positioning System

KPI Key Performance Indicator

NHTS National Household Travel Survey

TAZ Transport Analysis Zone

Apendix B - Household densities and income

Figure 11: Comparing eThekwini's population densities (a) and household incomes (b).

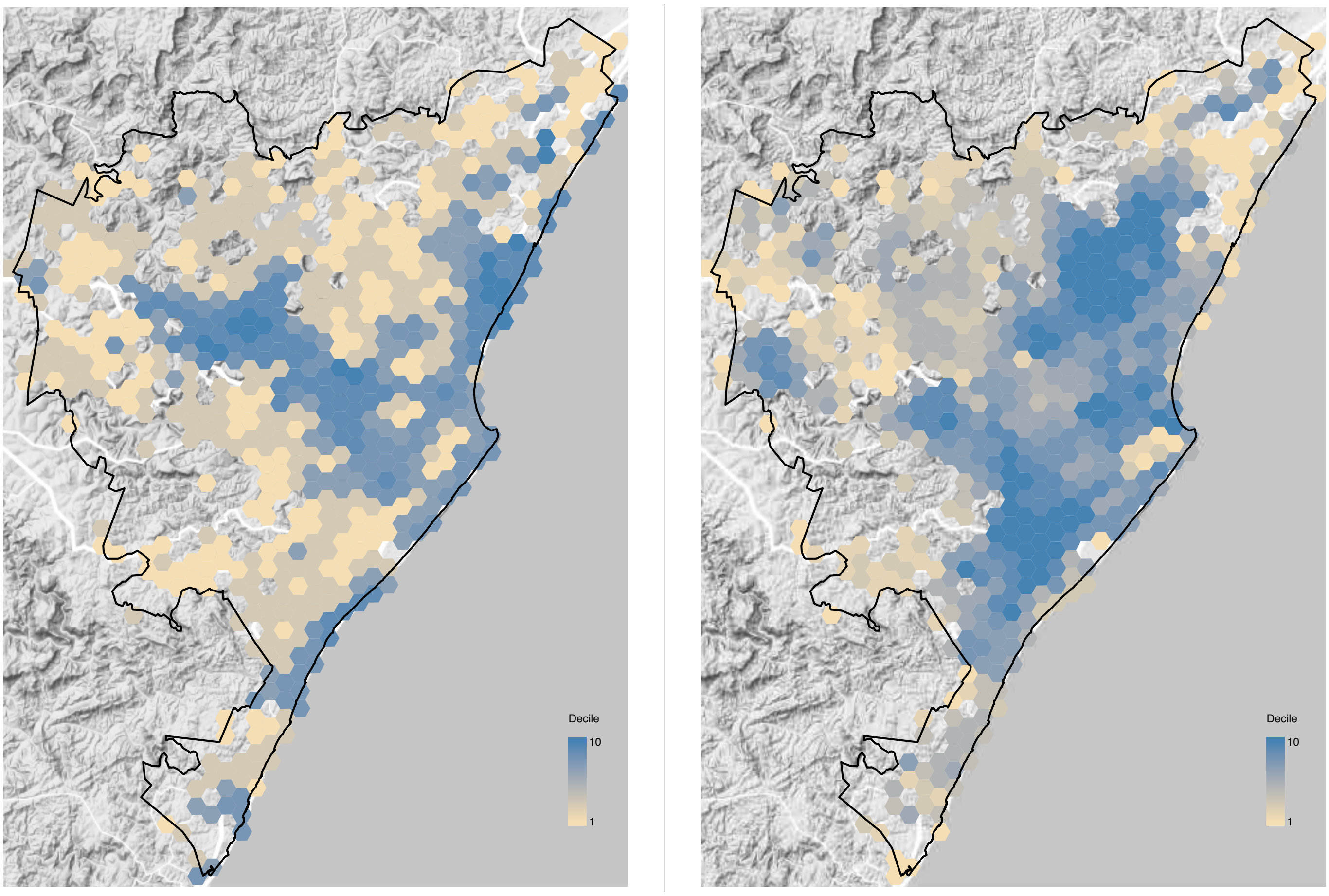
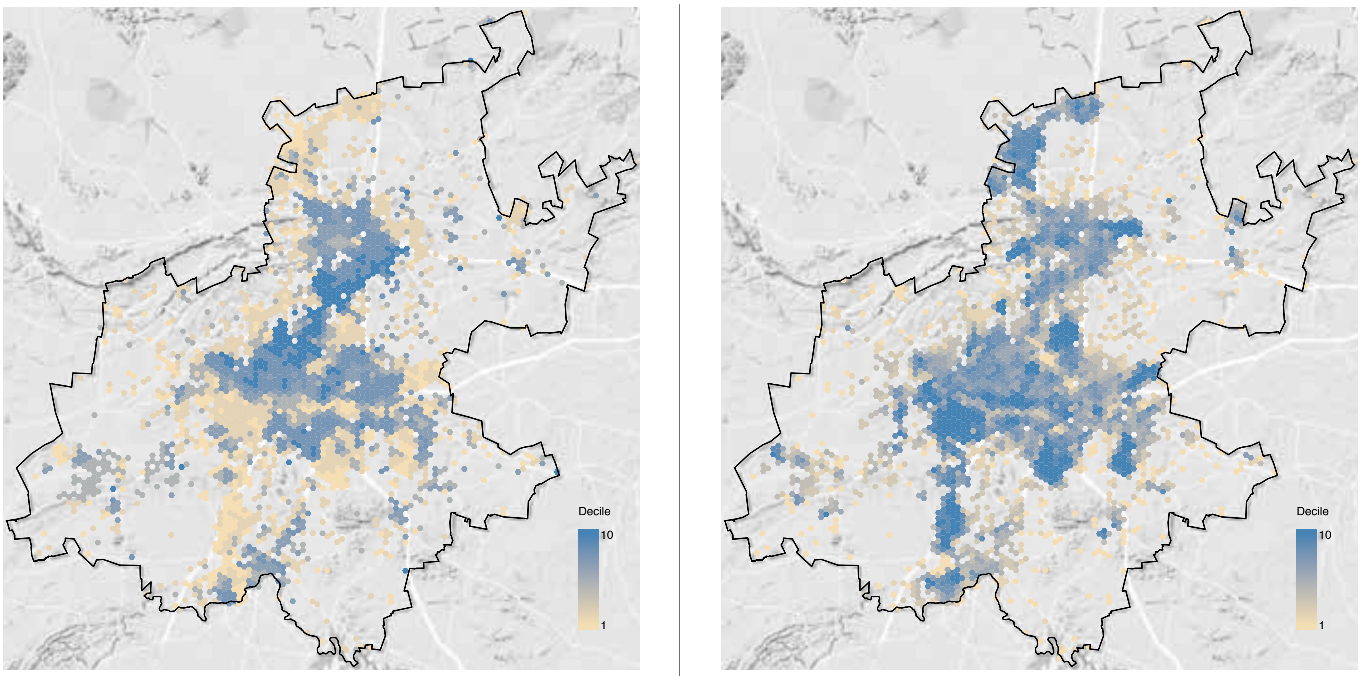


Figure 12: Comparing Gauteng's population densities (a) and household incomes (b).



Apendix C - Trip distributions

Figure 13: Trip distributions for eThekwini in the morning peak: a) outbound trips (generators); b) inbound trips (attractors).

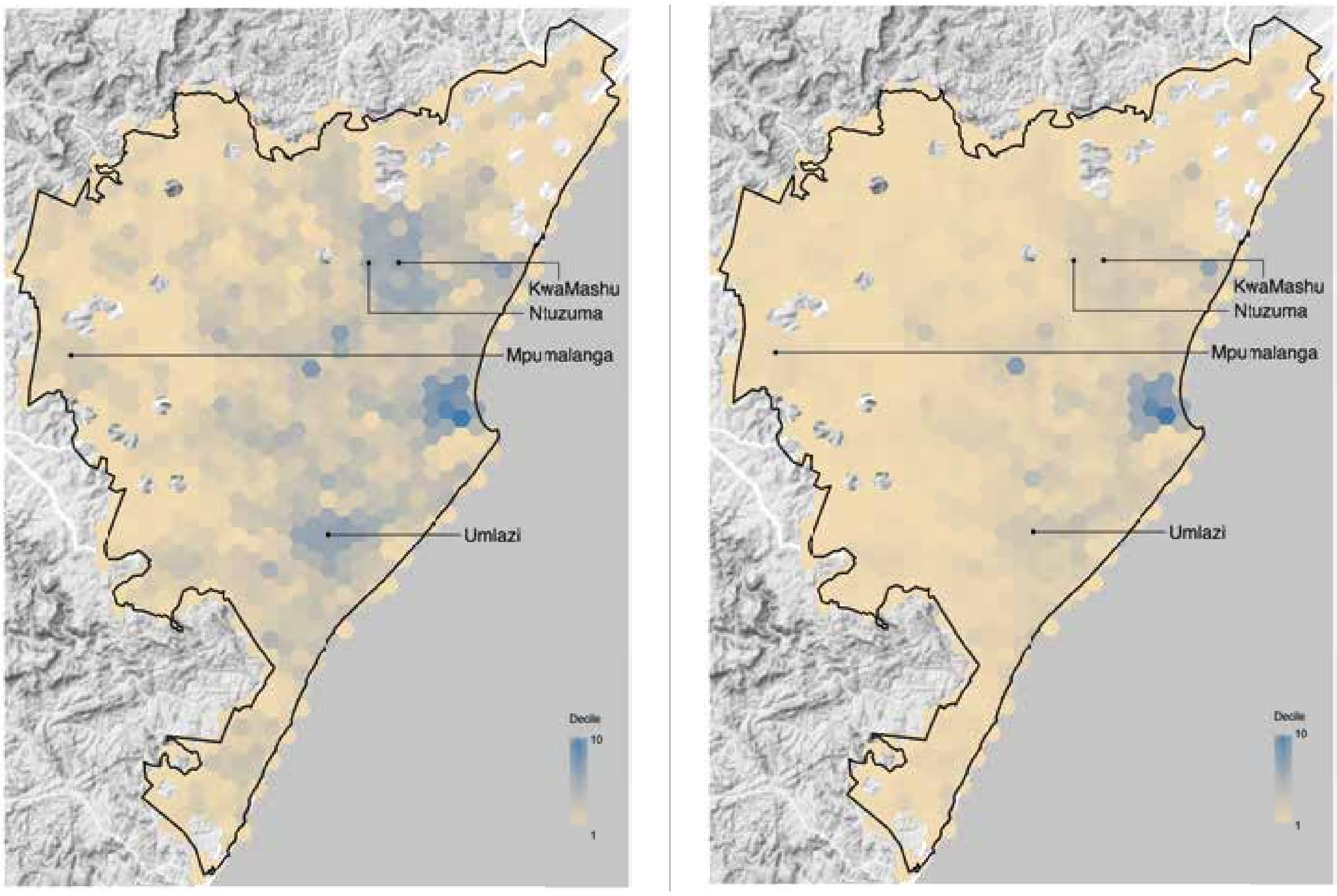
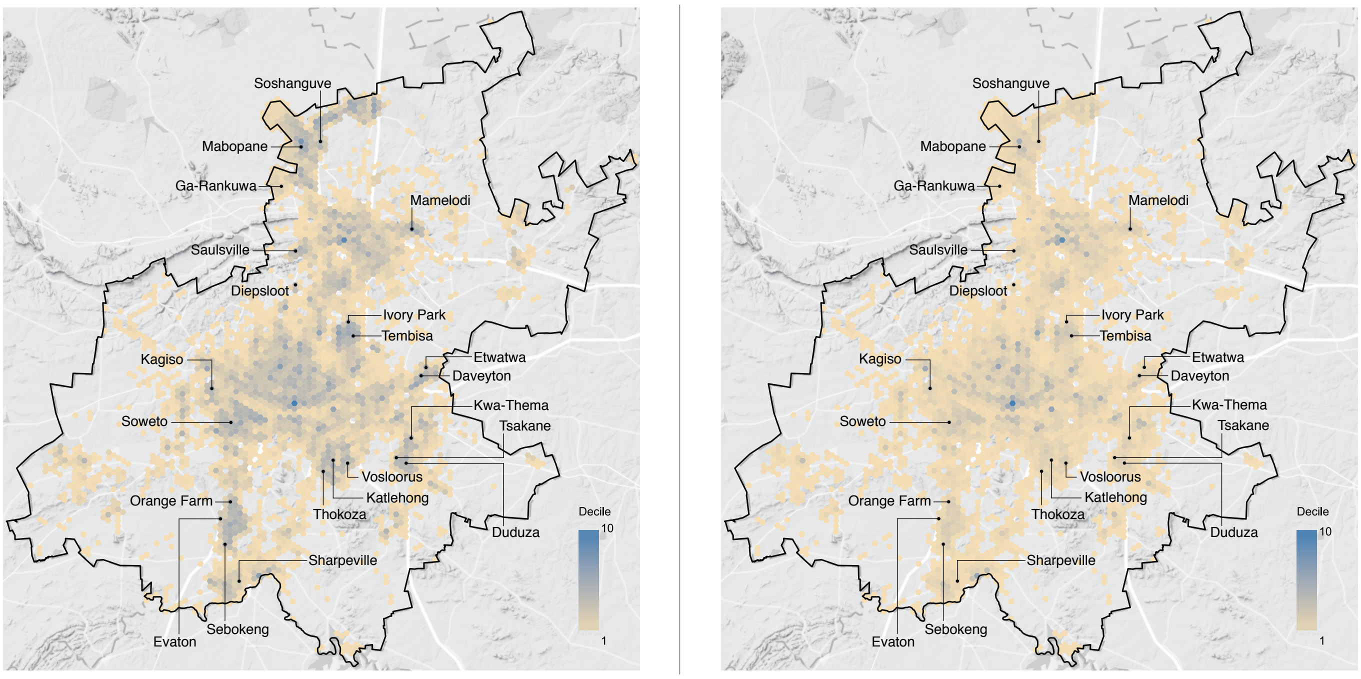


Figure 14: Trip distributions for Gauteng in the morning peak: a) outbound trips (generators); b) inbound trips (attractors).



Apendix D - Commercial closeness

Figure 15: Economic hotspots as a function of commercial vehicle activities in eThekwini.: a) commercial vehicle activities; b) closeness centrality.

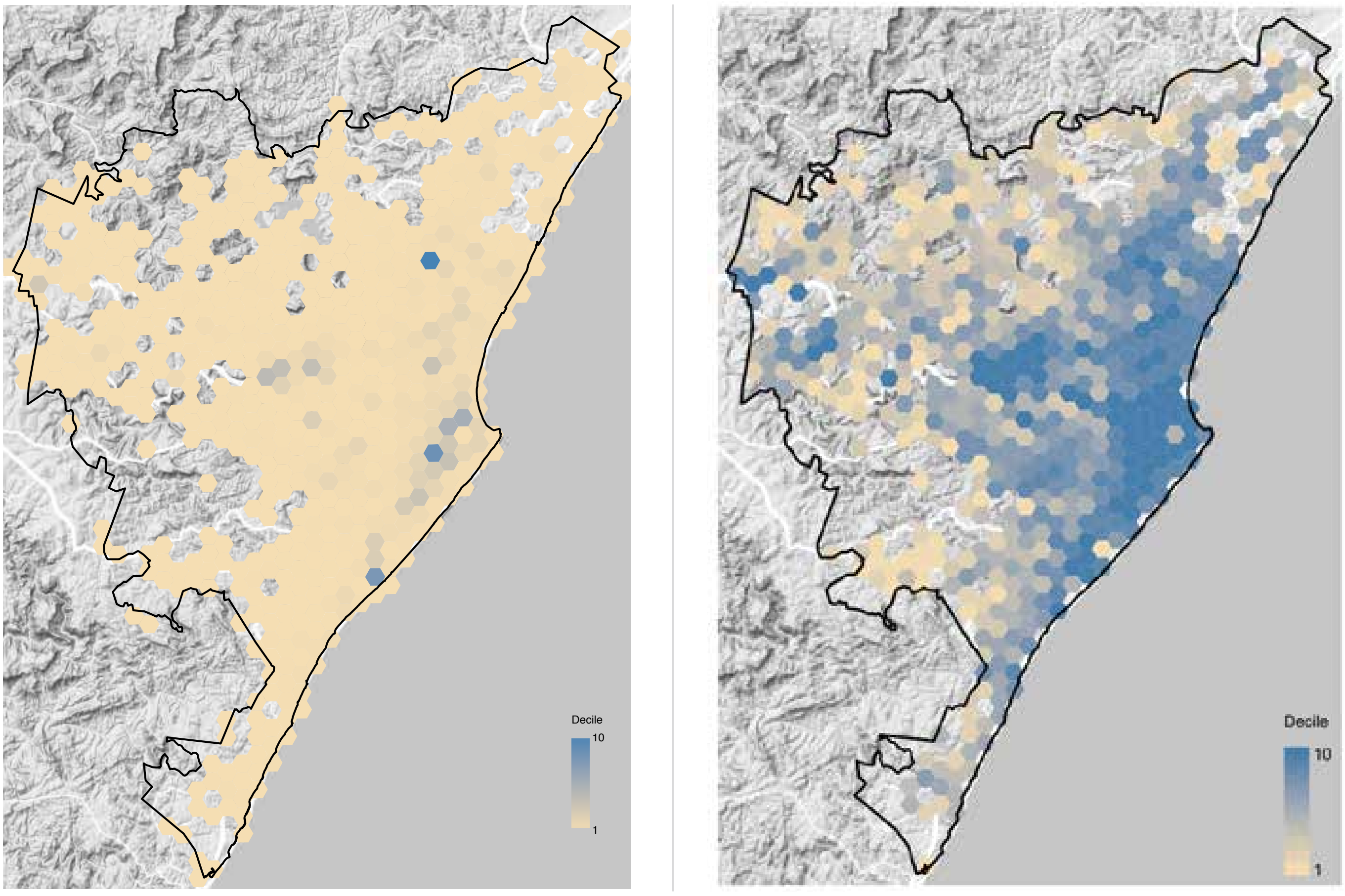
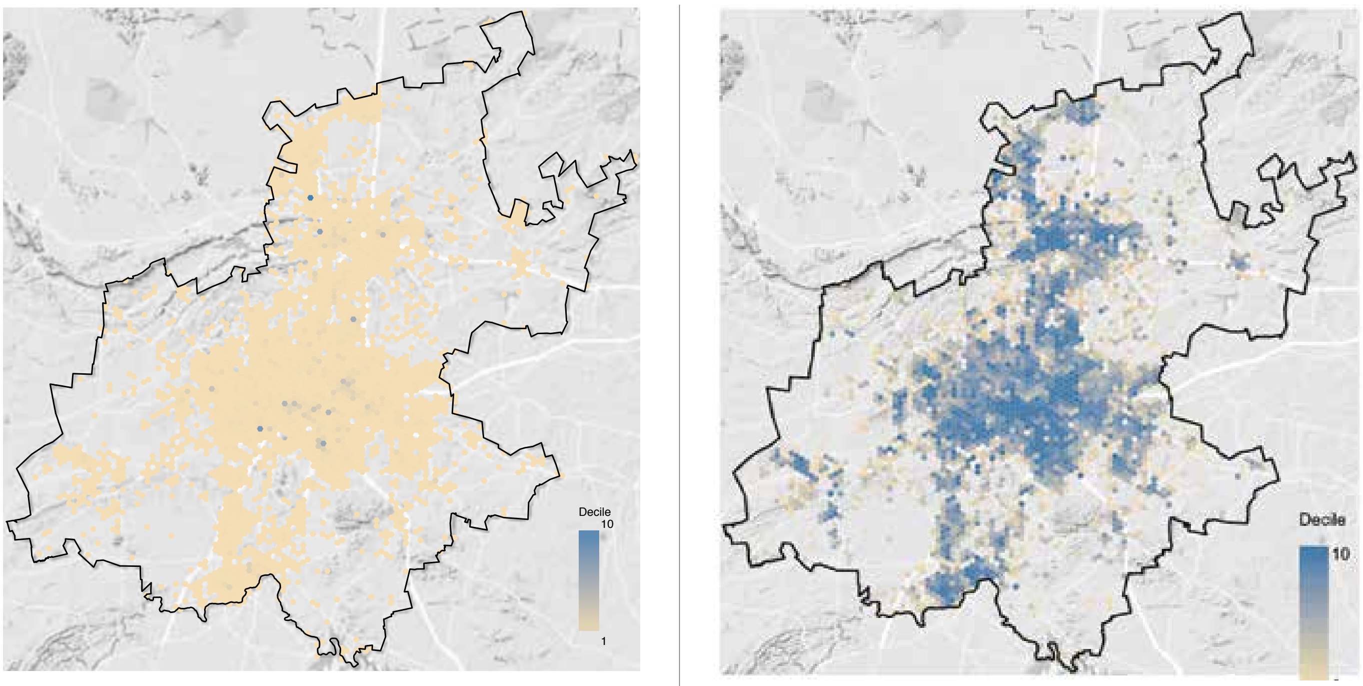


Figure 16: Economic hotspots as a function of commercial vehicle activities in Gauteng: a) commercial vehicle activities; b) closeness centrality.



Apendix E - Network structure

In this section we provide the final, prosed network structures for the different study areas.

Figure 17: Network version 2 structures for the City of Cape Town.

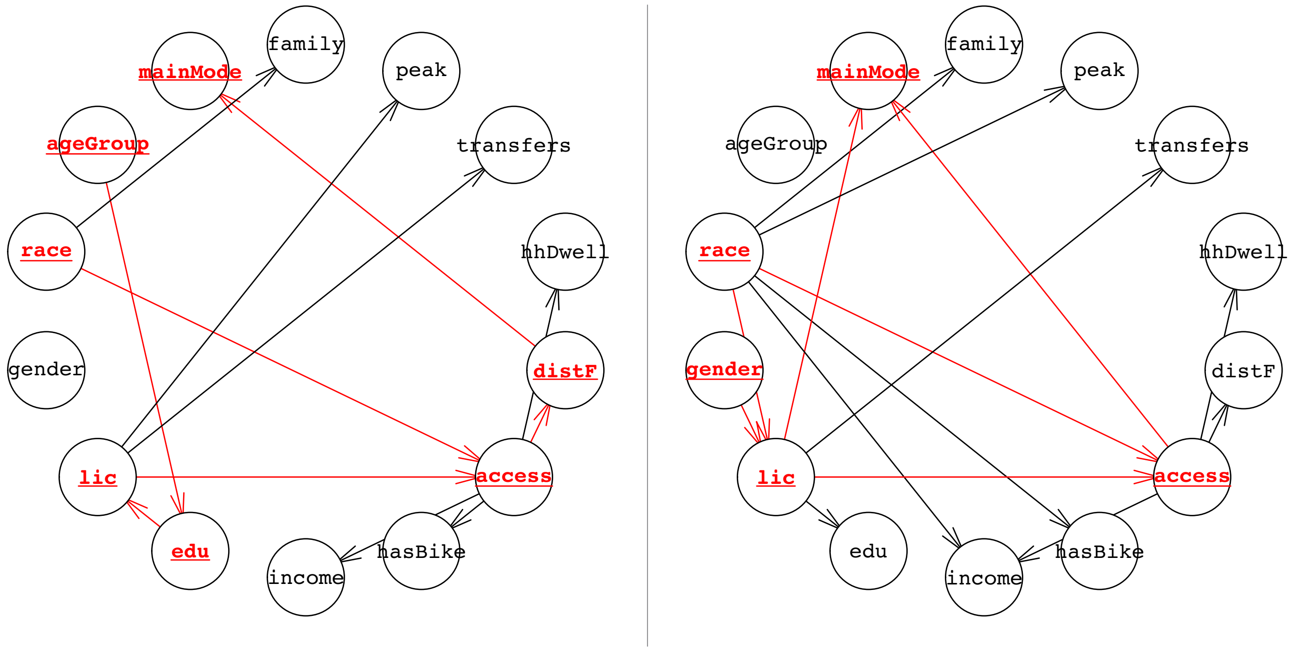


Figure 18: Network version 2 structures for eThekwini.

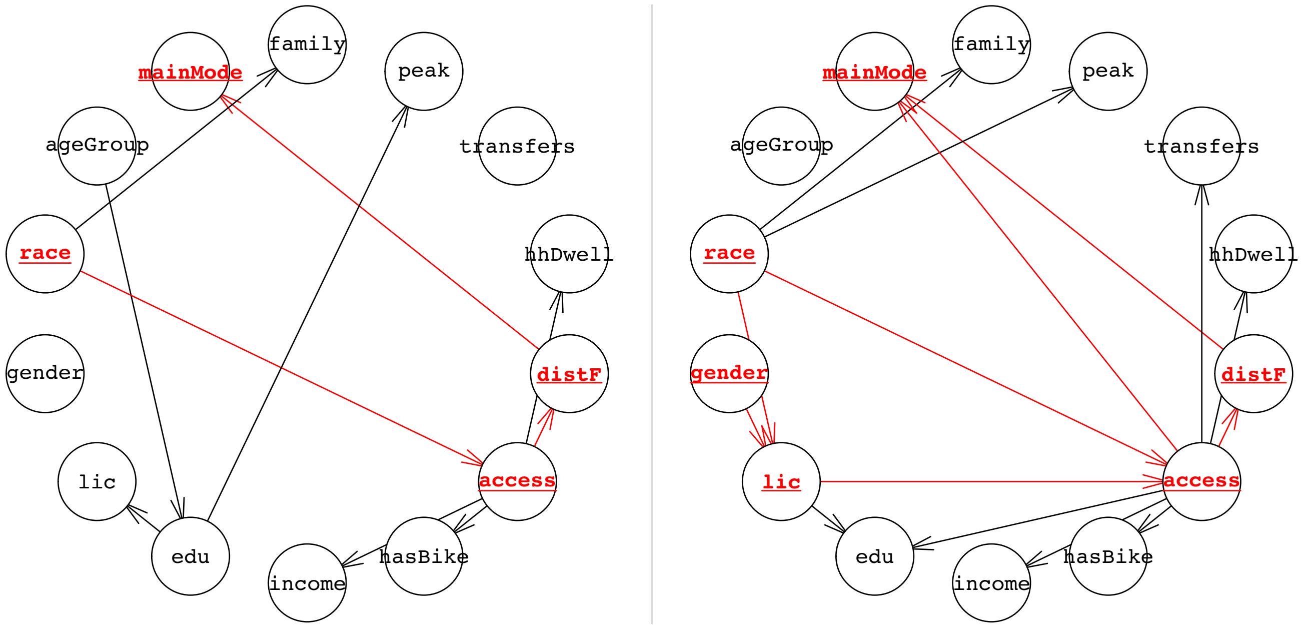


Figure 19: Network version 2 structures for Gauteng.

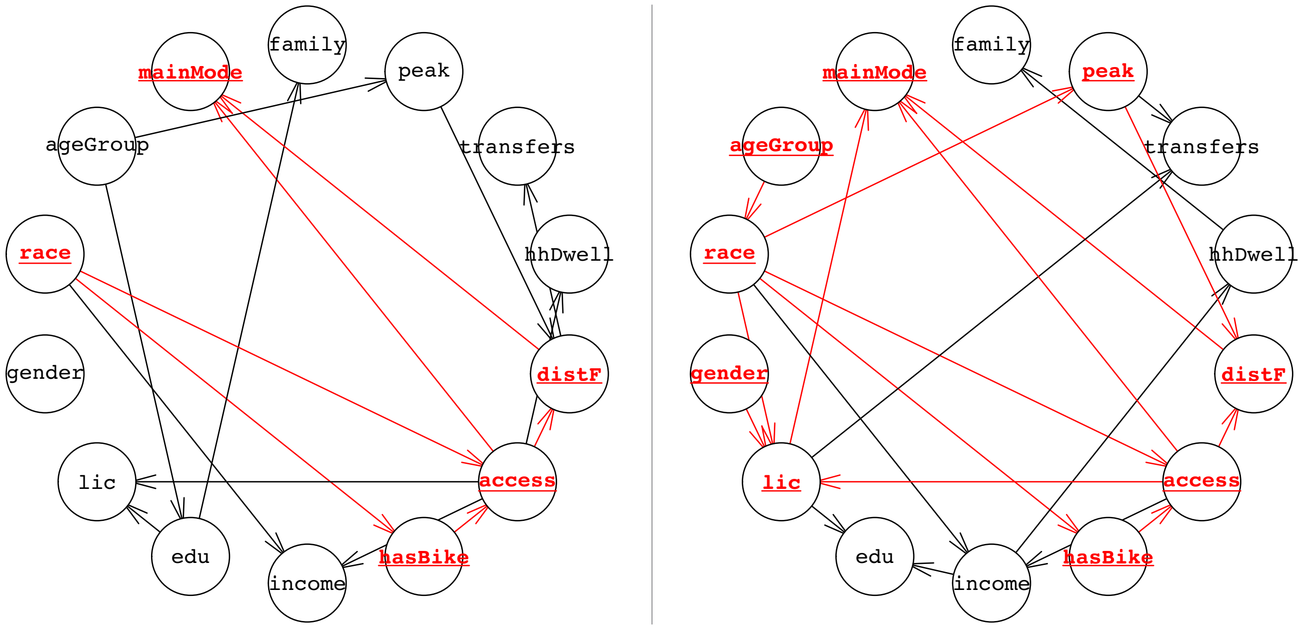
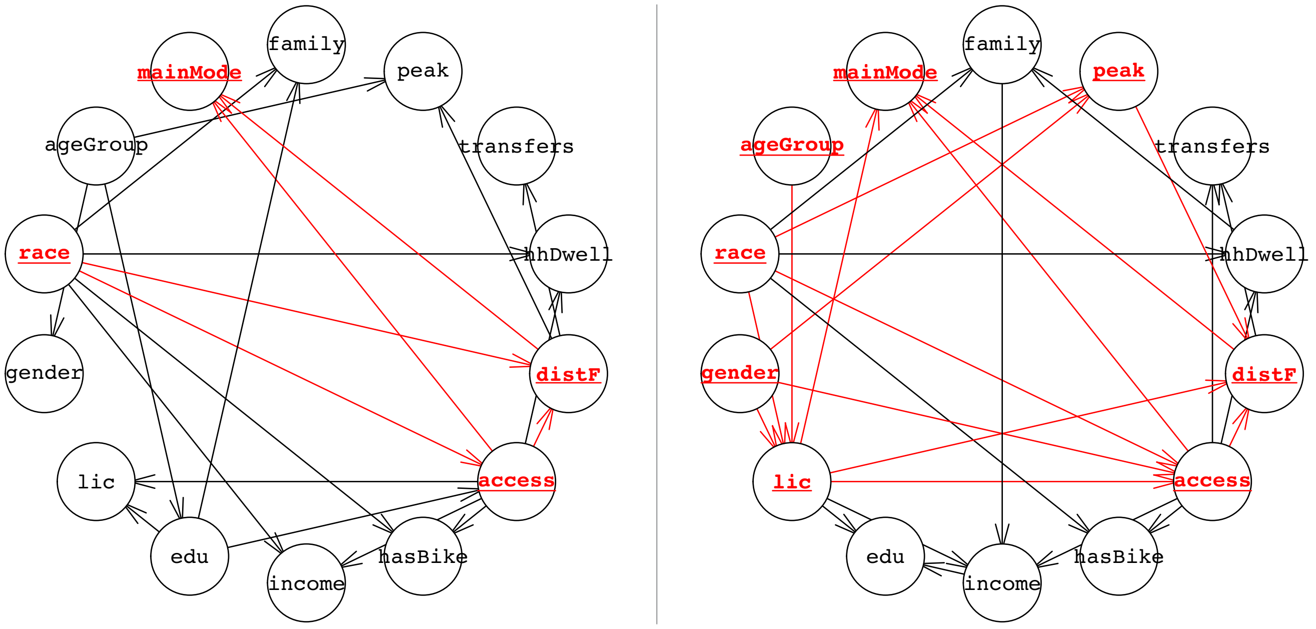


Figure 20: Network version 2 structures for the whole country.



This is one of nine background papers prepared as part of the South African Urbanisation Review (UR), which was commissioned by the Cities Support Programme (CSP) of the National Treasury to inform national policy and the Integrated Urban Development Framework (IUDF) on options for enhancing spatial integration. The other background papers and the overall report (`Managing Urbanisation to Achieve Inclusive Growth') is available from the CSP website.

NATIONAL TREASURY

Private Bag X115,
Pretoria, 0001
Tel: +27 12 315 5944

40 Church Square,
Pretoria, 0002
Fax: +27 12 406 9055

For further information :
Email: roland.hunter@treasury.gov.za
Visit: https://csp.treasury.gov.za
Tel: +27 (0)12 315 6515