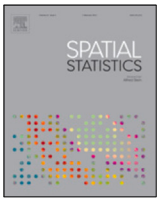
Spatial Statistics 55 (2023) 100748

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta



A Markov chain model for geographical accessibility

Renate N. Thiede [a,*], Inger N. Fabris-Rotelli [a], Pravesh Debba [b], Christopher W. Cleghorn [c]

[a] *Department of Statistics, University of Pretoria, Pretoria, Lynnwood Road, Pretoria, 0028, Gauteng, South Africa*
[b] *Council for Scientific and Industrial Research, Pretoria, South Africa*
[c] *School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa*

ARTICLE INFO

Article history:
Received 29 June 2022
Received in revised form 4 April 2023
Accepted 4 April 2023
Available online 11 April 2023

Keywords:
Accessibility
Markov chain
Spatial weights matrix
Linear network
Irregular lattice
Louvain clustering

ABSTRACT

Accessibility analyses are conducted for a variety of applications, including urban planning and public health studies. These applications may aggregate data at the level of administrative units, such as provinces or municipalities. Accessibility between administrative units can be quantified by travel distance. However, modelling the distances between all administrative units in a region is computationally expensive if a large number of administrative units is considered. We propose a methodology to model accessibility between administrative units as a homogeneous Markov chain, where the administrative units are states and standardised inverse travel distances act as transition probabilities. Single transitions are allowed only between adjacent administrative units, resulting in a sparse one-step transition probability matrix (TPM). Powers of the TPM are taken to obtain transition probabilities between non-adjacent units. The methodology assumes that the Markov property holds for travel between units. We apply the methodology to administrative units within Tshwane, South Africa, considering only major roads for the sake of computation. The results are compared to those obtained using Euclidean distance, and we show that using network distance yields more reasonable results. The proposed methodology is computationally efficient and can be used to estimate accessibility between any set of administrative units connected by a road network.

* Corresponding author.
E-mail addresses: renate.thiede@up.ac.za, renate.thiede@gmail.com (R.N. Thiede).

https://doi.org/10.1016/j.spasta.2023.100748

1. Introduction

Accessibility analyses measure the degree of access infrastructure provides to necessities such as healthcare, work and education opportunities, social interaction, and shops or markets (Mansour, 2016; Wigley et al., 2020; Netrdová and Nosek, 2020). Accessibility refers to ease of access based on factors such as travel time or distance, and can be thought of as the potential for travel. The need for accessibility research is highlighted by Sustainable Development Goal 9.[1] Accessibility between administrative units provides insights for applications where population data is spatially aggregated, such as infectious disease modelling, which typically reports infection counts at the level of administrative units. Disease spread is also often modelled between administrative units (Potgieter et al., 2021).

Modelling movement between administrative units becomes computationally expensive if a large number of administrative units is considered (Potgieter et al., 2021). The number of calculations increases with each additional unit added: at minimum, movement between m units requires $m!$ mobility values between units. On the other hand, measuring at a too low level of spatial aggregation leads to inaccurate results. Wigley et al. (2020) showed that accessibility results are highly dependent on the boundaries of administrative units. There is a need for a computationally efficient methodology that determines inter-administrative unit accessibility, and is easily scaled to different levels of spatial aggregation.

Accessibility analyses commonly conceptualise accessibility in terms of accessing specific utilities, such as healthcare facilities (Kelobonye et al., 2019; Mansour, 2016). In such cases, utilities are treated as the destinations or origins of travel. We however wish to consider the administrative units themselves as the origins and destinations, regardless of the location of utilities within those units, in order to quantify inter-administrative unit accessibility.

Our methodology provides a value that quantifies accessibility between every pair of administrative units in a region. These values are represented in a spatial weights matrix (SWM). SWMs assign a measure of spatial dependency between discrete spatial units (Stakhovych and Bijmolt, 2009), such as administrative units. For accessibility, the weights might represent travel distance, or time as in Netrdová and Nosek (2020).

We propose a statistical methodology for modelling accessibility between administrative units. Our approach is novel in its use of the relationship between SWMs and Markov chain transition probability matrices (TPMs), established by Bavaud (1998), to estimate accessibility. Markov chains have been used for movement modelling as far back as Brown (1970), who presents a comprehensive discussion on early work in this field. However, our scenario is not described therein. The states of the Markov chain are the administrative units, and row-standardised inverse travel distances between administrative units act as transition probabilities. Although transitions in such a Markov chain can include movement between non-adjacent units (Bavaud, 1998), we reduce computation by considering a single transition to be possible only between adjacent units. This assumption clearly holds for travel along road networks, and results in a sparse one-step TPM. Transitions between non-adjacent units are modelled as a series of transitions between adjacent units, obtainable from the n-step TPMs. For any $n > 1$, the n-step TPM is obtained as the sparse one-step TPM taken to the nth power. Numerically taking the limit of the n-step TPM as n tends to infinity gives the prominence index, which quantifies the relative accessibility of an administrative unit as a destination when unlimited transitions are allowed, and is independent of where the journey started. To the best of our knowledge, such an approach has not been used for accessibility modelling.

[1] https://unstats.un.org/sdgs/metadata/?Text=&Goal=9&Target=9.1

Estimation of the one-step matrix is the critical problem in this research. There is no one-size-fits-all approach or formal guidelines for constructing SWMs (Anselin, 2002). Rather, the choice of SWM must be appropriate to the problem and data. Our SWM makes use of the adjacency of administrative units and the distance between them. There are different ways to define distance, and the chosen definition has an impact on statistical estimation (Perret, 2011). Since people move mostly along the road network, we primarily consider network distance, i.e. the shortest distance between two points on the network (Yiu and Mamoulis, 2004). Our methodology reduces road network complexity in two ways. Firstly, it uses the contiguity of administrative units to simplify the storage and computation of the SWM as described above, and result in a sparse one-step TPM, making it computationally efficient to take the nth power in order to obtain n-step TPMs. Secondly, we do not calculate distance between each point on the road network, but rather between representative points in each administrative unit, obtained by clustering nodes on the road network via Louvain clustering (Blondel et al., 2008).

The rest of the paper proceeds as follows. Section 2 develops the one-step TPM construction methodology, as well as an alternative making use of Euclidean distance. Section 3 applies the methodology to administrative units within the City of Tshwane, South Africa, and compares the results to those using Euclidean distance. Section 4 discusses the proposed methodology and results, and Section 5 provides the conclusion.

2. Methodology

The primary outcome of the proposed methodology is a matrix which quantifies the accessibility between administrative units and their direct neighbours (the one-step TPM). Powers of this matrix are then taken to quantify the accessibility between units that are further apart, and to provide a measure of the overall accessibility of units as destinations. Section 2.1 provides the mathematical definitions required for the matrix construction methodology, and Section 2.2 describes the methodology.

2.1. Mathematical background

The administrative units are represented in an irregular spatial lattice. The relative locations of units in a lattice define their adjacency, and hence their neighbourhood. Adjacent units are considered neighbours, and a unit is considered a neighbour of itself. The road network is represented as a spatial linear network. A spatial linear network can be represented as a graph, where the lines are the edges, and the line endpoints and intersections are the nodes. The difference between a graph and a spatial linear network is that each node corresponds to a location in 2D space, and each edge corresponds to a line or arc in 2D space. Fig. 1(a) shows a region divided into administrative units, and (b) shows the road network represented as a spatial linear network using the `R` package `sfnetworks` (van der Meer et al., 2021).

Measuring distance on a network is not straightforward, since there may be multiple paths between two nodes on the network. We use network distance to refer to the shortest path between two nodes on the network.

Having explained how administrative units are represented and how distances can be calculated between them based on the road network, we now define a spatial weights matrix (SWM).

Definition 1 (*Spatial Weights Matrix (Bavaud, 1998)*)**.** Let $S = 1, 2, \ldots, m$ be a set of locations. Then the $m \times m$ matrix $P = [p_{ij}]$, $i, j = 1, \ldots, m$ is a spatial weights matrix if it meets the following conditions:

1. $p_{ij} \geq 0$,
2. $\sum_{i=1}^{m} p_{ij} = 1$ for $j = 1, \ldots, m$.

Each p_{ij} quantifies the strength of the spatial relationship in terms of probability between spatial regions i and j. SWMs are often defined to exclude self-neighbours (Stakhovych and Bijmolt, 2009), but the definition does not force this (Bavaud, 1998). Herein, we allow self-neighbours, such that

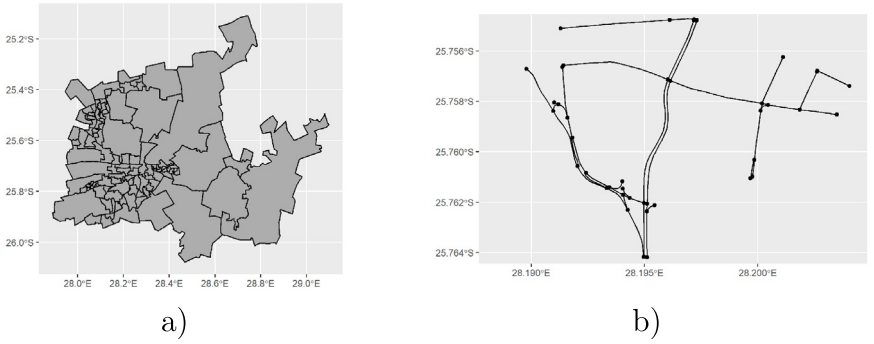


Fig. 1. Representation of the administrative units and roads. (a) An irregular lattice formed by subdividing a geographical region into administrative units. (b) A road network represented as a spatial linear network.

$p_{ii} \geq 0$. This is done to allow for journeys that remain within an administrative unit. Although this will in general lead to large diagonal values for low values of n, we believe that this allows for a realistic representation of how the population travels.

An SWM defined by Definition 1 satisfies the requirements for a matrix to be a Markov chain transition probability matrix (TPM), namely (1) non-negativity, i.e. $p_{ij} \geq 0$, and (2) row-standardised weights, i.e. $\sum_j p_{ij} = 1$ (Grimmett and Stirzaker, 2001, pg.215).

A Markov chain with finite states is ergodic if all its states are recurrent and aperiodic (Ross, 2007, pg.204). These conditions are satisfied if all the elements of P^n are greater than zero for some $n > 0$ (Bavaud, 1998). For an ergodic Markov chain, $P'\boldsymbol{\pi} = \boldsymbol{\pi}$ has a unique stationary distribution solution, $\pi_i \geq 0$, $\sum_i \pi_i = 1$. The term π_i takes on a specific interpretation in spatial modelling, namely the prominence index (Bavaud, 1998).

Definition 2 (*Prominence Index (Bavaud, 1998)*)**.** Let P be a spatial weights matrix that adheres to the requirements of a transition probability matrix of an ergodic Markov chain, and let $\boldsymbol{\pi}$ be the stationary distribution associated with P. Then the term π_i quantifies the total influence of unit i on the entire area under study, and is called the *prominence index* of unit i.

One of the necessary conditions for a finite Markov chain to be ergodic (and hence for the stationary distribution to exist) is irreducibility. A TPM will violate the irreducibility assumption if any state in the Markov chain is absorbing. We discuss how this could occur in our methodology in Section 2.2.

The final piece of mathematical background required for our methodology is network clustering. We obtain representative points within each unit by clustering nodes within the road network. This better represents the unit than a single point would, but is more computationally efficient than using all nodes in the network. We make use of Louvain clustering (Blondel et al., 2008), which is one of several network clustering techniques that group nodes in weighted networks based on topology (Emmons et al., 2016). Louvain clustering has been shown to outperform similar methods (Emmons et al., 2016), and scales well to large networks.

2.2. Proposed Markov chain accessibility model

The proposed model is a homogeneous Markov chain. The primary outcome here is a TPM $P = [p_{ij}]$ which quantifies accessibility between adjacent administrative units. Here the administrative units are the states of the Markov chain. Accessibility is quantified via the average inverse distance between two units, obtained by taking the reciprocal of the average network distance between the representative points within each unit. A transition between adjacent units is only possible if there exists a route between their representative points along the road network.

The value p_{ij} is the transition probability from unit i to unit j. It is calculated as the relative inverse travel distance from unit i to unit j, and represents the accessibility of unit j when starting in unit i. A value of p_{ij} that is close to 1 means that j is easily accessible from i, while a value close to 0 means that j is not easily accessible. If $p_{ij} = 0$, it means that unit j is not accessible from unit i. This will happen if the representative points of units i and j are not directly connected via the road network.

Fig. 2 gives an overview of the proposed methodology. The steps are described in detail in Appendix A. There are two inputs, namely the road network and the administrative unit boundaries. First, the road network is cleaned and converted to a spatial linear network using the `sfnetworks` package in R. Next, the road network is clipped by the unit boundaries to obtain the road network within a unit. Representative points are obtained in each administrative unit as follows. The network nodes within each unit are clustered via Louvain clustering. The centroid of each Louvain cluster is calculated in Euclidean space and projected onto the nearest network node. This network node is considered a representative node, which we refer to as a Louvain node. The distance associated with journeys remaining within a unit is taken as the average distance between the Louvain nodes within a unit. The unit boundaries are used to determine which units are adjacent, namely those with shared boundaries. The Louvain nodes of adjacent units are used to calculate the average distance between adjacent unit pairs. The average distances within and between units are then combined into a matrix of inverse distances. The one-step matrix P is obtained by applying row-wise normalisation to the matrix of inverse distances such that each value in the matrix is between 0 and 1, with higher values corresponding to higher levels of accessibility. The standardised inverse distances therefore act as probabilities. Each row of P adds up to 1. In the case where NAs are observed, p_{ij} is set to 0. A transition probability (inverse travel distance) of 0 implies a travel distance of infinity, i.e. it implies that the unit is inaccessible in 1 step.

The n-step TPMs are obtained by taking the nth power of the one-step TPM, $n > 1, n \in \mathbb{Z}$. For any given n, P^n gives the transition probabilities between units in a sequence of n steps between adjacent units. Taking $\lim_{n\to\infty} P^n$ will result in a matrix with identical rows, where each row gives the prominence index, as defined in Definition 2 in Section 2.1. Herein, $\lim_{n\to\infty} P^n$ is obtained by taking n to be sufficiently large. In this case, the prominence index of unit i can be interpreted as the probability of an infinite journey ending in unit i, regardless of its starting point.

The n-step matrices and the prominence index require the Markov chain to be irreducible, i.e. all states must be accessible in a finite number of transitions. The irreducibility assumption will be violated if an administrative unit i is not accessible from any of its neighbours (excluding itself). This will happen if the representative points of unit i are not connected to the representative points of any of its neighbours $j \neq i$ via the road network. A real-world system of administrative units and roads is unlikely to violate this assumption in practice, since administrative units will not typically be isolated from the road network. The risk lies in the way representative points are chosen. If we were to consider all the nodes within each unit, i.e. all intersections and line endpoints within a ward, as well as intersections between the road network and the unit boundaries, then each unit will be guaranteed to contain a node on the road network. However, calculating network distances between a large number of nodes becomes computationally prohibitive, especially at a low level of spatial aggregation. Some applications, such as Netrdová and Nosek (2020), use a single representative point per unit. Although this may still produce an accurate model, it has potentially serious consequences for the Markov chain. If the point chosen to represent a unit is not connected to the representative points of any of its first-order neighbours, the unit will be inaccessible from its neighbours. This motivates the use of multiple representative points in a unit, rather than a single point.

To determine if our use of the network distance is necessary, we introduce an alternative TPM construction, in which we use the Euclidean distance between the Louvain nodes. This may be obtained by a simple modification of the original construction. The detailed steps for constructing this alternative are provided in Appendix B.

Fig. 3 shows how distance between units is calculated for the original TPM construction and the Euclidean distance alternative, where distances are calculated along the purple lines. Distances are calculated along multiple routes in (a), namely the network distance between Louvain nodes. Euclidean distances are calculated between all the Louvain nodes in (b).

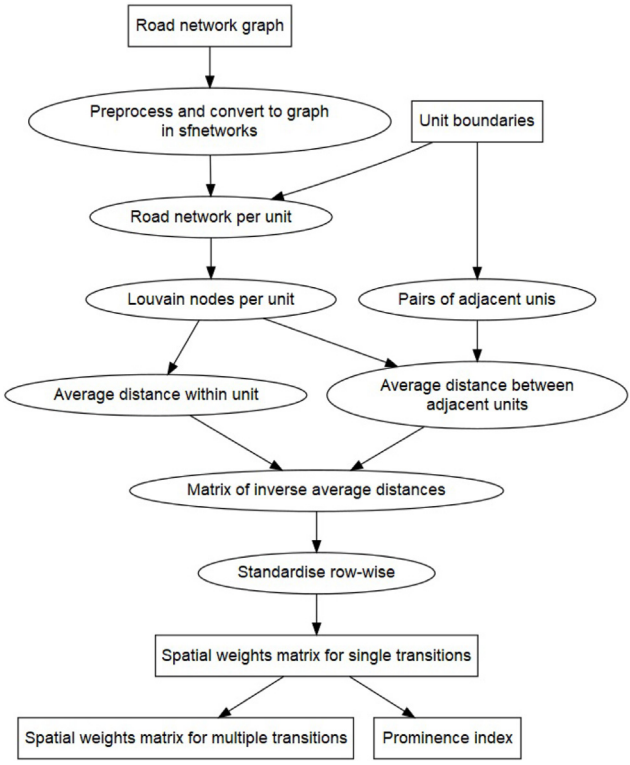


Fig. 2. Flowchart illustrating the process for obtaining the one-step TPM, and uses for this TPM. The inputs to the process are the road network graph, which is a spatial linear network, and the unit boundaries, where the units form an irregular lattice. Unit in this context refers to administrative units. The output of the process is a one-step TPM. The n-step TPMs are obtained by raising the one-step TPM to a finite power. The prominence index is obtained by raising the one-step TPM to the power n and obtaining the limit as n tends to infinity.

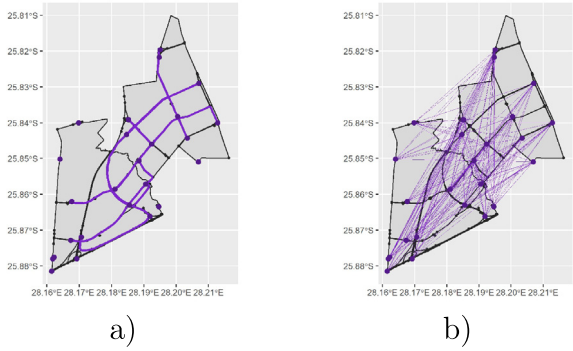


Fig. 3. Illustration of distance calculations. (a) The original construction, inverse network distance between Louvain nodes. (b) The alternative construction using inverse Euclidean distance between Louvain nodes.

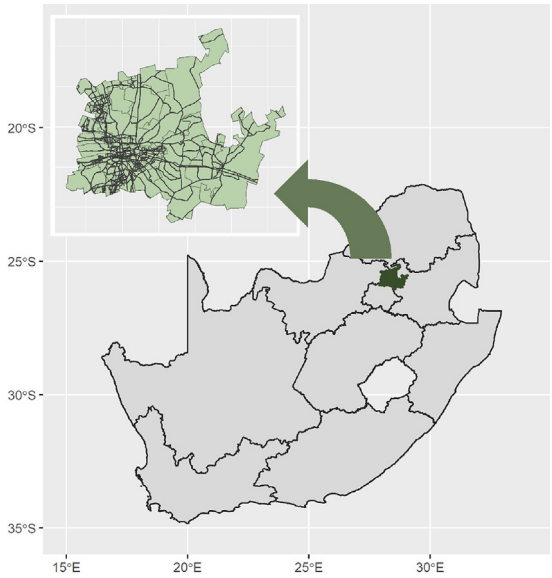


Fig. 4. The City of Tshwane municipality within Gauteng Province, South Africa. Inset: the road network of the City of Tshwane municipality.

The alternative has considerable advantages in terms of computation, since Euclidean distance is simpler to compute than network distance. However, since these distances are not along the road network, they may not accurately reflect travel distance, which could reduce the accuracy of the model.

3. Application

This section applies the proposed methodology to electoral wards within the City of Tshwane municipality, South Africa. We use our model to find the most accessible electoral wards from certain starting locations for a finite number of transitions, and to determine the prominence index. Finally, we compare the results of our method to a TPM construction using Euclidean distance instead of network distance.

3.1. Data

Fig. 4 shows the City of Tshwane municipality on the map of South Africa, with an inset showing the municipality and its road network. The municipality is divided into 107 electoral wards, and possesses a complex transport network with over 12 000 km of roads and paths, and over 3 000 km of major roads. The administrative boundary data was obtained from the Humanitarian Data Exchange. It was contributed by the OCHA Regional Office for Southern and Eastern Africa under the Creative Commons Attribution for Intergovernmental Organisations license.[2] The road network data was obtained from OpenStreetMap in August 2021. We did not consider the rail network or any other transport network, as most travel in South Africa occurs by road (Fabris-Rotelli et al.). For this analysis, we included motorways, primary, secondary and tertiary roads, and trunks. We excluded residential roads from the analysis for the sake of computation. The data did not include information on the directionality of the roads, therefore we represented the road network as an undirected graph. Since the included road types typically have two or more lanes, this is not expected to have a substantial effect on the results.

2 https://creativecommons.org/licenses/by/3.0/igo/legalcode

Table 1
Transition probabilities from the CBD.

n	W2	W3	W4
10	0.0005	0.0001	0.0002
20	0.0022	0.0005	0.001
50	0.0068	0.0041	0.0033
107	0.0094	0.0107	0.0057

3.2. Results from the proposed approach

The Markov chain accessibility model quantifies accessibility as the relative inverse distance between wards, expressed as a probability. An n-step TPM gives the probabilities of transitioning between wards in exactly n transitions between adjacent wards. The higher the transition probability from a ward i to a ward j for a given value of n, the more accessible j is from i in n transitions. Going forward, we will interchangeably refer to transition probability as accessibility.

As the number of transitions n tends to infinity, the transition probabilities converge to the prominence index. The prominence index of ward i gives the transition probability to i for a potentially infinite journey, and does not depend on a particular origin.

For a finite number of transitions, we obtain results with respect to four origin wards, shown in Fig. 5. Ward W1 covers part of Tshwane's central business district (CBD) and is connected to its neighbours by a complex network of major roads. Wards W2 and W3, in contrast, are both located in townships towards the edge of the city. Ward W2 is in the Soshanguve–Mabopane area in northwestern Tshwane, and Ward W3 is in the Mamelodi area in eastern Tshwane. Ward W4 is located in central Centurion, which includes high-value economic areas and borders on another township.

Fig. 6 shows the top 10 most accessible wards when starting from ward W1 for $n = 10$, $n = 20$, $n = 50$ and $n = 107$ transitions, and the prominence index. Ward W1 is outlined. As n increases, i.e. as possible journeys cross more wards, some outlying wards become more accessible. By $n = 107$, the results are starting to look similar to the prominence index in (e).

Fig. 7 shows the top 10 most accessible wards when starting from wards W2, W3 and W4, for $n = 107$. Table 1 gives the transition probabilities to the CBD for these origins as n increases.

We now consider the results for the alternative TPM construction using Euclidean distance. Results for finite transitions ($n < \infty$) are presented only with respect to ward W1 and for $n = 10$. Fig. 8(a) shows the 10 most accessible wards for the alternative. The difference is clear between this and the network distance-based construction in Fig. 6(a). The network distance-based construction consider the large wards to the west of W1 to be more accessible, based on the major roads passing through them, while the Euclidean distance-based construction does not.

Fig. 8(b) shows the top 10 most accessible wards assuming an infinite journey from any starting point, i.e. the wards with the highest prominence index. As with the finite transitions, we see a clear difference between the network distance-based and Euclidean distance-based methods regarding the large western wards. The original identifies a large western ward as being very prominent, and identifies two southern wards around the central Centurion area (ward W4 and a neighbour), which are connected to the CBD via major highways. The alternative identifies a groups of prominent wards in northwestern Tshwane around Soshanguve and Mabopane (ward W2), and does not pick up the large western wards.

4. Discussion

4.1. Proposed Markov chain accessibility model

Using a ward in the CBD as starting point, we illustrated how the relative accessibility of wards changed as the number of transitions increased (Fig. 6). Wards close to the origin were the most accessible for a low number of transitions. Some outlying wards became more accessible for a higher

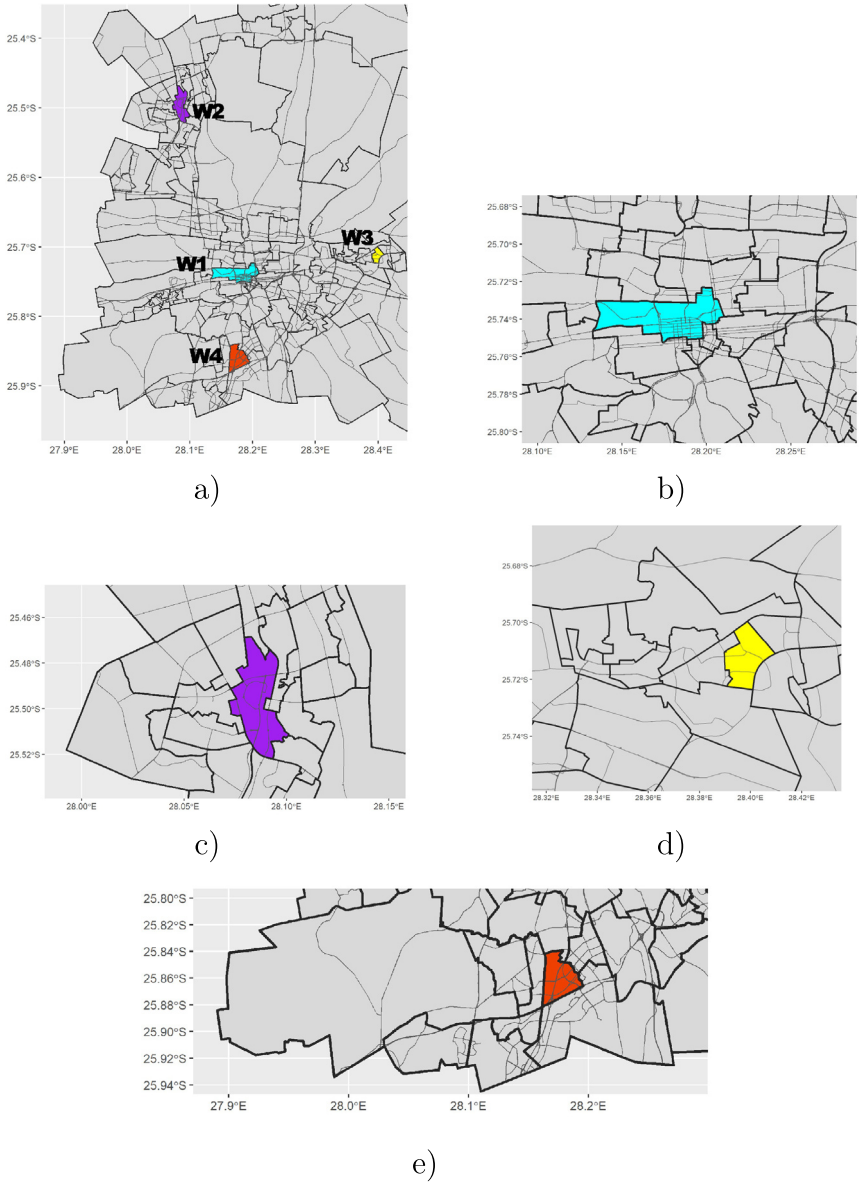


Fig. 5. Wards W1-4. (a) The wards within their surroundings: ward W1 is denoted in blue, W2 in purple, W3 in yellow and W4 in orange. (b) Ward W2 covers part of Soshanguve and Mabopane, townships in northwestern Tshwane. (c) Ward W1 covers part of the CBD of Tshwane, and contains a complex network of major roads. (d) Ward W3 covers part of Mamelodi, a township in eastern Tshwane. (e) Ward W4 is located in central Centurion, which includes high-value economic areas and borders on a township. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

number of transitions. These wards correspond to residential or economical areas that have contact with the CBD, and the results of the model suggest that this contact is represented by the road network. Even for a high number of transitions, some wards nearer to the CBD remain among the most accessible.

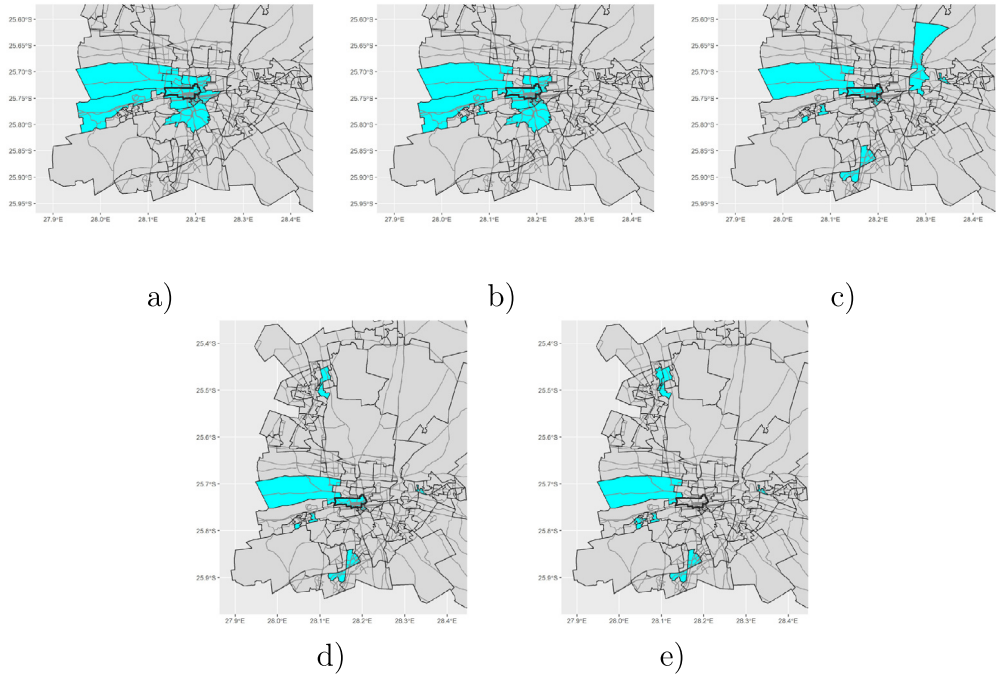


Fig. 6. Top 10 accessible wards starting from ward W1, in the Tshwane CBD, for increasing values of n. (a) Most accessible wards for $n = 10$. (b) Most accessible wards for $n = 20$. (c) Most accessible wards for $n = 50$. (d) Most accessible wards for $n = 107$. (e) The prominence index, i.e. most accessible wards in the limit as n approaches infinity.

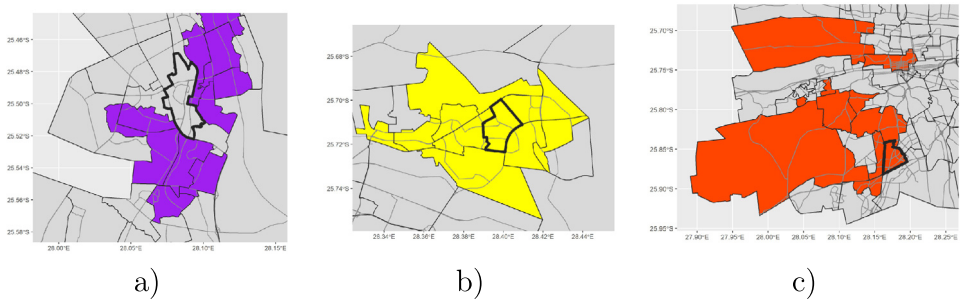


Fig. 7. Top 10 accessible wards for $n = 107$, starting from ward W2 (Soshanguve/Mabopane) in (a), W3 (Mamelodi) in (b) and W4 (Centurion) in (c). Note that in (a), the origin ward itself (outlined) is not among the top 10 most accessible from the origin.

The prominence index identified the top 10 most accessible wards in the municipality, for journeys of unlimited length regardless of origin ward. These wards include a ward in western Tshwane, which is connected to other wards by a major highway, and some small wards toward the edges of the city which all overlap with township areas. The prominence index also identifies two wards in Centurion, which includes economic and residential areas, and borders on another township, though the township itself is not included in these wards. The accessibility of these areas makes sense in the context of Tshwane, and shows interaction between the locations of where people live, and how the transport system has developed. The township locations were identified by

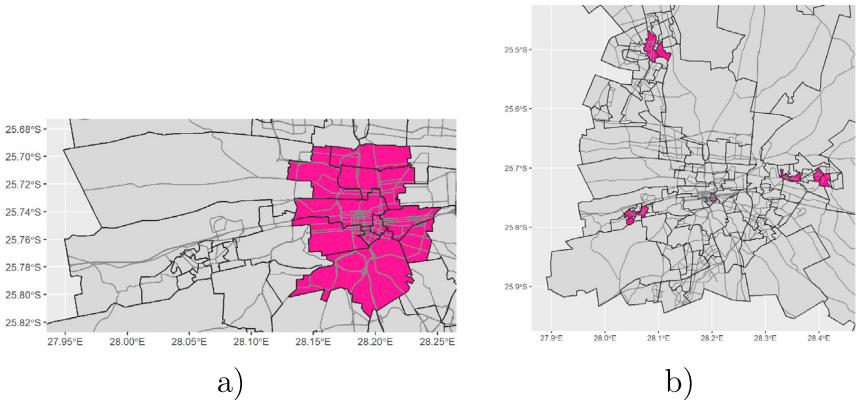


Fig. 8. Results for the Euclidean distance-based TPM construction. (a) Results for $n = 10$ transitions, starting from ward W1. (b) The prominence index, i.e. results for n in the limit, regardless of origin.

the previous government, and the initial residents had no choice but to live where the government dictated. However, these locations have remained popular nearly 30 years after the end of forced resettlement, partly for historical reasons, but also because these areas are convenient in terms of transport, since transport infrastructure developed to service these areas.

It is interesting to note that the prominence index identifies wards in townships as among the most accessible. Given the levels of inequality within the municipality, one might expect that more privileged areas would enjoy better accessibility. A possible explanation is that wealthier citizens prefer to reside further away from major roads due to noise, and rely on quieter residential roads for access.

We also investigated the accessibility of wards when starting from wards in Soshanguve/Mabopane (ward W2), Mamelodi (ward W3) and Centurion (ward W4). The top 10 wards for ward W2 (Fig. 7(a)) and ward W3 (b) are clustered around their respective origins. This implies that their accessibility to more distant areas, including those with more work opportunities, does not necessarily improve with increasing n. This may be due to the fact that they are in residential areas, and not directly connected to the CBD by major roads. The top 10 wards for ward W4 (c) include mostly residential and semi-rural wards to the east of ward W4, which houses many of the residents working in W4, and wards in and around the CBD (ward W1). This suggests that ward W4 is better connected to wards that are further away, allowing for easier travel to e.g. the CBD. These results indicate that the wards in Soshanguve/Mabopane and Mamelodi, i.e. townships, do not enjoy the same levels of accessibility to the CBD as the ward in Centurion, an area with a higher level of development and more major roads. Although the township wards, W2 and W3, were highly accessible from the CBD (ward W1), the reverse is not the case. We do however see that accessibility to the CBD increases with n, as shown in Table 1. This illustrates the validity of the model, while highlighting the accessibility challenges faced by residents in wards W3 and W2.

To justify our use of network distance, we investigated an alternative TPM construction obtained by using Euclidean distance. Although this alternative presents a considerable computational advantage, it produced less intuitive results, in line with our expectations.

We found that all wards within the City of Tshwane were accessible from each other given enough transitions, i.e. sufficiently large n. This is because all wards are intersected by major roads. While this is expected in a well developed, mostly urban municipality, this may not be the case for all municipalities, especially in rural areas. In the event that a ward is not intersected by a major road, one may remove it from the study area, or force it to be accessible by connecting an arbitrary point (such as the centroid) to a major road in a neighbouring ward by a straight line. If multiple wards in a study area are unreachable, however, this suggests that the type of road chosen is not appropriate, and that more road types should be included (e.g. residential roads or tracks).

4.2. Limitations and future work

The methodology has a number of limitations. Firstly, the assumption of the Markov property may be too strong in some cases, for example if a unit k is heavily used as a route from some unit i to j. In this case, the probability of moving from k to j may depend heavily on whether one has arrived in k from i. Secondly, the methodology does not account for the real ability of people to travel along the road network. The methodology assumes that if a road exists in a ward, everyone within the ward is able to travel along the road in equal measure. However, given the diversity and poverty in South Africa and the heterogeneous condition of roads, this may not be the case. This data is not easy to come by, but if available could be incorporated in future research. Thirdly, sensitivity to the clustering method was not investigated. Louvain clustering was chosen for its computational advantage, but is by no means the only viable clustering method. A full sensitivity analysis merits a future study.

In future work, the model could be adapted in a number of ways. Additional road types could be included, such as residential roads, and travel time could be considered instead of distance. This would require the use of different TPMs for different times of day and days of the week, as travel time varies with traffic conditions. This is a departure from the homogeneous Markov chain proposed herein, which is based on fixed inverse travel distances. Information on traffic volume could be incorporated into the TPM construction, as well as the directionality of roads. Directionality information could influence the results in regions that contain many one-way roads. Incorporating information on bus and taxi routes could also add valuable insights, although this is complicated for South Africa, as the publishing of taxi route data is a sensitive political issue. The model could also be run for different administrative levels. South Africa is divided into administrative areas at various levels, such as municipalities and provinces. Modelling at a finer level, such as wards, will deliver insights into accessibility over shorter local journeys, while a TPM at municipality level will be more suited to investigate accessibility across the country. The only caution here is that the model should not be run in conditions where the administrative units are mutually adjacent such that the one-step TPM is not sparse. This would increase the computational cost of obtaining n-step TPMs and counteracts the advantage of our method.

A relevant future application of this research is to measure hospital accessibility in order to reduce maternal deaths. Wigley et al. (2020) showed that over half of sub-Saharan African countries do not provide sufficient access for women of childbearing age to appropriate healthcare facilities. This indicates a need for further research to identify at risk areas.

Mathematical properties of the TPM could be investigated, such as the rate of convergence to the limiting distribution. The probabilities in the TPM could also be used to explore the relationships between administrative units, for instance to determine whether it tends to group administrative units such that there is poor accessibility between groups.

Finally, the results of the model should be validated against existing accessibility models. This is not straightforward: the new method is not directly comparable to existing accessibility models, as these typically quantify accessibility between utilities, rather than between discrete areas like administrative units. Such comparisons merit a future study, for which the planning is already underway.

5. Conclusion

This paper developed a computationally simple statistical approach to model accessibility between administrative units based on the road network, using the link between spatial weight matrices and Markov chain transition probability matrices presented by Bavaud (1998). The sparse one-step TPM drastically reduces the computational cost of determining accessibility. The methodology is adaptable, allowing for alternative choices of representative points and distance or time measures. The methodology was applied to the set of electoral wards in the City of Tshwane municipality, South Africa. Based on network distance between Louvain nodes, we were able to identify wards that were accessible from Tshwane's CBD as well as other wards within the municipality, and used the prominence index to find wards that were the most accessible regardless

of starting location. These results align with our expectations based on conditions on the ground. Lastly, we compared the use of network distance to Euclidean distance, and illustrated our reasons for preferring the construction based on network distance despite the computational cost. The proposed methodology is computationally efficient and flexible and can be used to estimate accessibility between any set of administrative units connected by a road network.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to acknowledge Rüdiger Thiede for his contributions to the data preprocessing approach. This work is based on research supported in part by the National Research Foundation of South Africa (Grant Number 137785 and CoE-MaSS ref #2022-018-MAC-Road) and the NRF-SASA Academic Statistics Grant. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF.

Appendix A. Original TPM construction

Denote the spatial lattice of the entire study region by $S_R = \{U_1, \ldots, U_N\}$, where N is the number of spatial units in S_R.

1. Obtain the adjacency matrix $A = [a_{ij}]$ of S_R, where $a_{ij} = 1$ if units U_i and U_j are adjacent, and 0 otherwise.
2. Let $D = [d_{ij}]$ be the distance matrix of S_R, where d_{ij} represents the distance between U_i and U_j as calculated below. Initialise $d_{ij} = 0$.
3. For $i = 1, \ldots, N$, obtain the set of Louvain nodes $L_i \neq \emptyset$ within each unit U_i:
    - 3.1. Perform Louvain clustering on U_i to identify the Louvain clusters within that unit, $C_{i1}, C_{i2}, \ldots, C_{in_i}$, where n_i is the number of Louvain clusters in U_i.
    - 3.2. For $l = 1, \ldots, n_i$, obtain each Louvain node in L_i:
        - 3.2.1. Obtain the centroid x_{il} of C_{il} in Euclidean space.
        - 3.2.2. Project x_{il} onto the nearest network node, where nearness is defined by Euclidean distance. This network node, denoted by λ_{il}, is a Louvain node and is added to the set L_i.
4. For $i = 1, \ldots, N$, obtain the average network distance between unit U_i and each of its first-order neighbours, based on their Louvain nodes:
    - 4.1. Obtain the set of first-order neighbours B_i of U_i from the adjacency matrix A, i.e. $B_i = \{U_j : a_{ij} = 1\}$. Note that $U_i \in B_i$.
    - 4.2. For $U_j \in B_i$, calculate $\bar{A}_N(L_i, L_j)$, the average network distance between L_i and L_j:
        - 4.2.1. Calculate the network distance between the Louvain nodes within U_i and each of its neighbours, i.e. $d_N(\lambda_{il}, \lambda_{jk})$, $l = 1, \ldots, n_i$, $k = 1, \ldots, n_j$.
        - 4.2.2. Calculate the average of these distances

$$\bar{A}_N(L_i, L_j) = \frac{1}{n_i n_j} \sum_{l=1}^{n_i} \sum_{k=1}^{n_j} d_N(\lambda_{il}, \lambda_{jk}) \text{ for } i \neq j$$

and

$$\bar{A}_N(L_i, L_i) = \frac{1}{n_i^2 - 1} \sum_{l=1}^{n_i} \sum_{k=1}^{n_i} d_N(\lambda_{il}, \lambda_{ik}),$$

since $d_N(\lambda_{il}, \lambda_{il}) = 0$.

4.3. Set $d_{ij} = \bar{A}_N(L_i, L_j)$.

5. Let $V = [v_{ij}] = [\frac{1}{d_{ij}}]$, i.e. the element-wise reciprocal of D. This gives the inverse distances. Where $d_{ij} = 0$, set $v_{ij} = \text{NA}$.
6. The TPM, P, is then the row-wise standardisation of V, i.e.

$$P = [p_{ij}] = \left[\frac{v_{ij}}{\sum_{m=1}^{N} v_{im}}\right] = \left[\frac{\frac{1}{d_{ij}}}{\sum_{m=1}^{N} \frac{1}{d_{im}}}\right],$$

with $p_{ij} = \text{NA}$ where $d_{ij} = 0$.

Appendix B. Alternative TPM construction

The alternative construction considers the Euclidean distance between Louvain nodes. Modify Step 4.2.1 of the original TPM construction as follows:

4.2.1. Calculate the Euclidean distance between the Louvain nodes within U_i and each of its neighbours, i.e. $d_E(\lambda_{il}, \lambda_{jk})$, $l = 1, \ldots, n_i$, $k = 1, \ldots, n_j$, with $l \neq k$ if $i = j$.

References

Anselin, L., 2002. Under the hood issues in the specification and interpretation of spatial regression models. Agric. Econ. 27 (3), 247–267.

Bavaud, F., 1998. Models for spatial weights: a systematic look. Geogr. Anal. 30 (2), 153–171.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008 (10), P10008.

Brown, L.A., 1970. On the use of Markov chains in movement research. Econ. Geogr. 46 (sup1), 393–403.

Emmons, S., Kobourov, S., Gallant, M., Börner, K., 2016. Analysis of network clustering algorithms and cluster quality metrics at scale. PLoS One 11 (7), e0159161.

Fabris-Rotelli, I., Holloway, J., Kimmie, Z., Archibald, S., Debba, P., Manjoo-Docrat, R., le Roux, A., Dudeni-Tlhone, N., Janse van Rensburg, C., Thiede, R., Abdelatif, N., Makhanya, S., Potgieter, A., 2022. A Spatial SEIR model for COVID-19 in South Africa, 2, 14–45.

Grimmett, G., Stirzaker, D., 2001. Probability and Random Processes. Oxford University Press, p. 215.

Kelobonye, K., McCarney, G., Xia, J.C., Swapan, M.S.H., Mao, F., Zhou, H., 2019. Relative accessibility analysis for key land uses: A spatial equity perspective. J. Transp. Geogr. 75, 82–93.

Mansour, S., 2016. Spatial analysis of public health facilities in Riyadh Governorate, Saudi Arabia: A GIS-based study to assess geographic variations of service Provision and accessibility. Geo-Spatial Inf. Sci. 19 (1), 26–38.

Netrdová, P., Nosek, V., 2020. Spatial dimension of unemployment: Space-time analysis using real-time accessibility in Czechia. ISPRS Int. J. Geo-Inf. 9 (6), 401.

Perret, J.K., 2011. A proposal for an alternative spatial weight matrix under consideration of the distribution of economic activity. Technical Report, Schumpeter Discussion Papers.

Potgieter, A., Fabris-Rotelli, I.N., Kimmie, Z., Dudeni-Tlhone, N., Holloway, J., Janse van Rensburg, C., Thiede, R.N., Debba, P., Docrat, R., Abdelatif, N., Khuluse-Makhanya, S., 2021. Modelling representative population mobility for COVID-19 spatial transmission in South Africa. Front. Big Data 4.

Ross, S.M., 2007. Introduction to Probability Models. Academic Press, p. 204.

Stakhovych, S., Bijmolt, T.H., 2009. Specification of spatial models: A simulation study on weights matrices. Pap. Reg. Sci. 88 (2), 389–408.

van der Meer, L., Abad, L., Gilardi, A., Lovelace, R., 2021. Sfnetworks: Tidy geospatial networks. URL https://CRAN.R-project.org/package=sfnetworks. R package version 0.5.2.

Wigley, A., Tejedor-Garavito, N., Alegana, V., Carioli, A., Ruktanonchai, C.W., Pezzulo, C., Matthews, Z., Tatem, A., Nilsen, K., 2020. Measuring the availability and geographical accessibility of maternal health services across sub-Saharan Africa. BMC Med. 18 (1), 1–10.

Yiu, M.L., Mamoulis, N., 2004. Clustering objects on a spatial network. In: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data. pp. 443–454.