Modelling representative population mobility for COVID-19 spatial transmission in South Africa

A. Potgieter [1], I. N. Fabris-Rotelli [1,*], Z. Kimmie [2], N. Dudeni-Tlhone [3], J. Holloway [3], C. Janse van Rensburg [4], R. Thiede [1], P. Debba [3,5], R. Manjoo-Docrat [5], N. Abdelatif [4], and S. Makhanya [6]

[1] *University of Pretoria, Department of Statistics, South Africa*
[2] *Foundation of Human Rights, South Africa*
[3] *Council for Scientific and Industrial Research, South Africa*
[4] *Biostatistics Research Unit, South African Medical Research Council*
[5] *Department of Statistics and Actuarial Science, University of Witwatersrand, South Africa*
[6] *IBM, South Africa*

Correspondence*:
Inger Fabris-Rotelli
inger.fabris-rotelli@up.ac.za

ABSTRACT

The COVID-19 pandemic starting in the first half of 2020 has changed the lives of everyone across the world. Reduced mobility was essential due to it being the largest impact possible against the spread of the little understood SARS-CoV-2 virus. To understand the spread, a comprehension of human mobility patterns is needed. The use of mobility data in modelling is thus essential to capture the intrinsic spread through the population. It is necessary to determine to what extent mobility data convey the same message of mobility within a region. This paper compares different mobility data sources by constructing spatial weight matrices and further compares the results through hierarchical clustering. This provides insight for the user into which data provides what type of information and in what situations a particular source is most useful.

Keywords: COVID-19, spatial, mobility, spatial weight matrices, principal component analysis, hierarchical clustering

1 INTRODUCTION

The COVID-19 pandemic starting in the first half of 2020 has changed the lives of everyone across the world. From working from home at all hours, using less public and personal transport, home-schooling under lock down, to economic strife and anxiety; predicting such changes would have been near impossible a priori. By far the largest impact, aside from the economic troubles many find themselves in, is reduced mobility. Daily commuting has been much reduced due to various lockdown measures across countries internationally. In addition, international flights and cross border travel was restricted for significant periods of time, even between regions in some countries.

Reduced mobility was essential, however, due to it being the largest impact possible against the spread of the little understood SARS-CoV-2 virus. Social distancing and stay at home instructions were understood and implemented internationally. These instructions were seen as the best protection for the individual, as well as being the means to prevent overload on the hospital systems, which would otherwise result in inflated death rates. These protection mechanisms are formed on an understanding of the basic nature of the spatial spread of the virus. A virus spreads via a host, whom it relies on to move amongst other susceptibles. The more movement and interaction performed by the host, the more the virus is able to spread. It is thus imperative to incorporate a spatial element when modelling the spread of the COVID-19 pandemic.

Quantifying mobility patterns of people facilitates a more accurate understanding of the spread of the disease. An individual's ability to physically "lock down" and stay at home was affected by economic inequality, as shown in a US study [1]. In South Africa, this economic inequality is extreme, with the World Bank recognising South Africa, in 2019, as having the worst inequality in the world[1].

While the strict lockdown introduced by the South African government from 27 March 2020 delayed the first wave, the mobility was by no means completely reduced due to many living day-to-day for food. Food parcel queues from food donations were a large focus during the first half of the pandemic in South Africa, as the risk of contracting COVID-19 was overridden by the need for food. Such queues, and the use of public transport during these times, heightened the transmission risk of COVID-19 in South Africa, even while lockdown rules were in place. A full lockdown was therefore not possible, and spatial interaction continued between individuals from different regions within South Africa. Modelling regions in isolation will therefore not capture the influence of this mobility on the spread of COVID-19 in South Africa.

The use of mobility data in modelling is thus essential to capture the intrinsic spread through the population. A common source is mobile phone location data, which has been utilized previously for epidemiological modelling [2, 3, 4, 5, 6, 7]. However, this data is difficult to obtain due to increasing privacy concerns. In addition, there are most often a number of network providers in a region, each with certain market share. Without data access from all, or at least, the largest providers, representativeness and mobile phone penetration will be limited and should be used with caution.

It is necessary to determine to what extent different sources of mobility data convey the same message of mobility within a region. In this paper we demonstrate, through the use of principal component analysis as well as hierarchical clustering, how different sources of spatial mobility data at various resolutions can lead to different conclusions with regards to spatial unit connectivity. Spatial connectivity is an essential first step in spatial modeling, providing a quantification of the spatial dependency between spatial units. Herein, we compare the use of different spatial weight matrices in quantifying how different spatial units relate. We also discuss the advantages of different sources and how they can be harnessed when modelling the spread of a virus. We do this by using principal component analysis in order to condense the information that can be gained from a spatial weight matrix and then using hierarchical clustering to identify the strongest spatial associations and to essentially put on display what type of relationships the spatial weight matrix is identifying. This is to the best of our knowledge the first time this exact combination has been used for this purpose.

The mobility data available for South Africa is presented in Section 2. The methodology for constructing mobility matrices is developed in Section 3, and the results are presented in Section 4. Section 5 discusses the results and Section 6 concludes.

[1] See the following link

2 DATA

Mobility data are often used to understand various issues ranging from epidemic modelling, transport planning and management, communication network improvement and urban planning [8, 9]. Asgari et al [8] indicates that mobility goes far beyond mere geographical movement of humans, but provides a comprehensive perspective on human interactions that could be considered from spatial, temporal, and contextual aspects. Human mobility is one of the aspects of mobility that gained attention from the global spread of infectious diseases as with the recent COVID-19 pandemic. A variety of technologies including navigation sensors, wireless technologies, and cellular communication technologies are used to position humans in space [10]. A study by [9] provides a comprehensive overview of the different types of human mobility patterns data. These include those data types that capture both the wider (city-wide) and minute (building-wide or large structure) human movements, for example, cellular services records (CSRs), surrounding WiFi access point records (SWAPRs), Global Positioning System locations (GPSLs), geotagged social media (GTSM), public transport smart card records (PTSCRs), bluetooth detection records (BDRs), and WiFi probe request records (WFPRs). The analysis methods range from data visualisation to statistical analysis methods (classification and clustering), heuristic logic, graph theory and optimization techniques.

The administrative divisions of South Africa are summarised in Table 1. In order of increasing spatial resolution these are country, province, district municipality, local municipality, and ward, labelled as administrative levels 0 through 4 respectively. To facilitate the comparison of different sources of spatial information, it is first necessary to aggregate the data from each source to the same spatial resolution. Increasing the resolution of spatial data can be achieved through methods such as small area estimation or spatial micro-simulation (see e.g. [11, 12]). These methods are somewhat involved and require the use of auxiliary information or assumptions that are unlikely to be true. In this paper we investigate aggregating down to the lowest spatial resolution used by our data sources. While this is relatively straightforward to accomplish, it potentially results in the loss of information.

Table 1. South Africa administrative boundaries

Administrative level	Spatial unit name	Number of spatial units
0	Country	1
1	Province	9
2	District municipality	52
3	Local municipality	213
4	Ward	4392

Two mobility data types were available for this research. The first is freely available data shared by Facebook, and the second is mobility data made available by a South African cellular provider for the COVID-19 response in 2020.

2.1 Facebook Data for Good

Multiple geographically indexed datasets have been made freely available for use by Facebook through their “Facebook data for good” initiative. These datasets serve to aid researchers and policymakers in understanding the spread of COVID-19 [2].

This paper utilises one of these available datasets, namely the “Movement range maps” dataset. The data indicates the change in mobility, $F_i^{(t)} \in (-1, 1)$ (as a percentage), for a spatial unit i on a given day t over the period 1 March 2020 – 28 February 2021 relative to a one-week baseline calculated in February 2020. The daily values for each district municipality were calculated by determining the number of so-called “Bing tiles”[3] that each inhabitant visited on a given day (place of residence being determined by the location where users most often spend their nights). After incorporating some degree of noise, the average number of tiles visited by the inhabitants was determined and expressed relative to the baseline. The full description of how these values were calculated is available in the Appendix. The spatial resolution for units of this data are district municipalities, thus the data is at administrative level 2.

Figure 1 shows the data for all district municipalities, with the average across the district municipalities shown in red. The figure demonstrates that the average mobility nationally dropped significantly in late March. This corresponds to when South Africa entered its first hard lockdown on the 27th of March 2020 (see Table 2). The hard lockdown imposed severe restrictions on travel and constituted a strict stay at home directive. Only essential workers were allowed to leave their homes. Furthermore, the average change in mobility is primarily negative over the entire study period, indicating that mobility patterns remain more constrained than before the hard lockdown. The first COVID-19 case was discovered on 5 March 2020 and the lockdown announcement was made a week later on 15 March. This could explain the drop in mobility already seen from early March.

Notable benefits of using this data are that the data is freely available and could potentially act as a very representative proxy for human mobility, as Facebook services are not constrained to specific mobile network providers. In addition, all the cellular network providers in South Africa provide a free version of Facebook called Facebook Zero. Even though it is known that not all South Africans have a Facebook account, the Facebook mobility data may provide an acceptable level of representativety for mobility within the country. It is also clear that a large amount of the original data was censored in order to preserve user privacy and thus the data is at a sparse level of spatial resolution (administrative level 2). The data is also not specific with regards to the direction of spatial mobility. Daily observations only indicate whether individuals were more or less mobile and do not indicate the places towards which this mobility was directed.

2.2 Mobile network data

The growing popularity and widespread use of mobile devices has led to massive amounts of data being produced at any given point in time all around the world. Mobile phone data can be collected either passively by mobile services providers or through the use of mobile applications. The ease with which such large quantities of data can be gathered makes cellular data attractive for researchers.

Mobile phone data has been used numerous times in the field of spatial epidemiology to model the spread of various diseases, including cholera [2, 3], dengue [4, 5] and malaria [4, 5]. Following the outbreak of the COVID-19 pandemic, the governments of various countries across the world began collecting cellular device user data in an attempt to aid the conception and implementation of non-pharmaceutical

[2] See the following link

[3] See Bing tiles

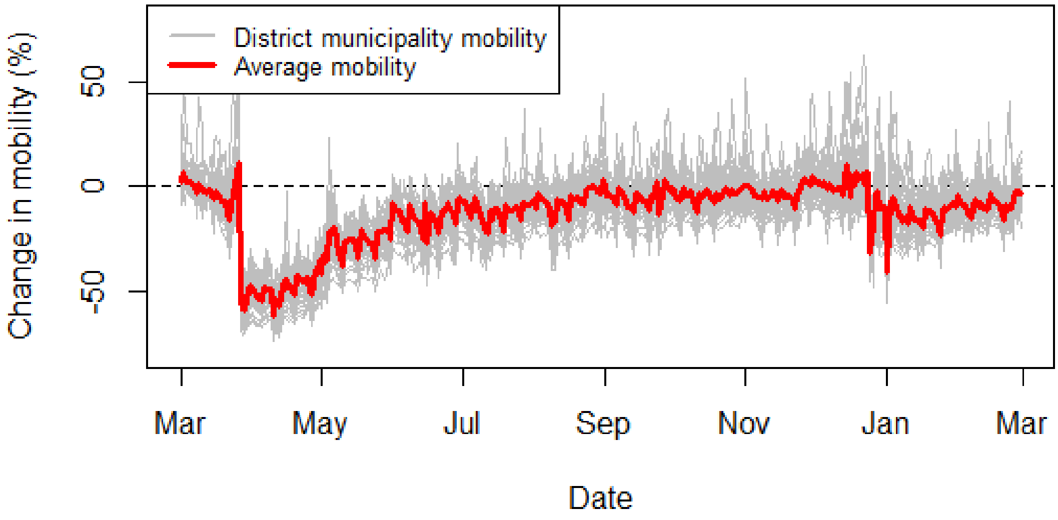

Figure 1. Facebook for good data (1 March 2020 – 28 February 2021)

interventions [13, 14, 15, 16]. This data has since been used by researchers to clearly establish a correlation between population mobility and COVID-19 case numbers [16, 17, 18, 19].

Limitations of mobile phone data exist. First and foremost of these is the issue of user privacy. Mobile phone data could potentially be misused to identify specific individuals and thus cellular providers are often hesitant to provide researchers with such data [15, 20]. Such data is often aggregated to a low spatial resolution to prevent this as well as reduce noise, but this comes at the cost of some data specificity. Another potential drawback of mobile phone data is high computational cost. Due to very high mobile phone penetration rates, mobile phone data may consist of a number of entries in the order of billions. The computational cost of processing such datasets is prohibitive, potentially preventing analysis.

For this paper, anonymised mobile phone data was obtained from a local mobile network provider. In South Africa, the mobile phone penetration level is estimated to be as high as 95%[4]. The provider utilised in this paper is one of the largest providers in the country, with an estimated market share of 42%.

The data indicates the number of mobile phone users $m_{ij}^{(t)}$ that travelled to ward j from ward i on day t for the period 2 March - 12 May 2020. The data is at administrative level 4, which is the highest spatial resolution reasonably possible while preserving some level of privacy of exact user location. To compare insights gained from this data and the previously discussed Facebook dataset, it would first be necessary to aggregate the mobile phone data to the same spatial resolution which is administrative level 2. In South Africa, each ward has a unique 8-digit ID code. The first three digits of this code indicates the district municipality that the ward is a part of. For example, the ward ID 9344007 indicates that the ward is part of the district municipality with code 934, namely Vhembe. In order to aggregate the data to district municipality level, one could replace the ward IDs of the observations with their district municipality codes (i.e. only the first 3 digits), whereupon rows with identical origin and destination codes would be discarded.

[4] See the following link 1, link 2

The mobile phone data at administrative level 2 is thus given by

$$M_{I,J}^{(t)} = \sum_{i \in I, j \in J} m_{ij}^{(t)},$$

where I and J are district municipalities and i and j are wards as previously indicated. Transitions contained within a single district municipality are thus discarded. Initial analysis revealed that this caused an average of 26% of daily observations to be discarded. The retained data is displayed in Figure 2. The representation differs to that of Figure 1 as the data provides transitions between regions in this case. We once again notice a sharp decline in population mobility in late March.

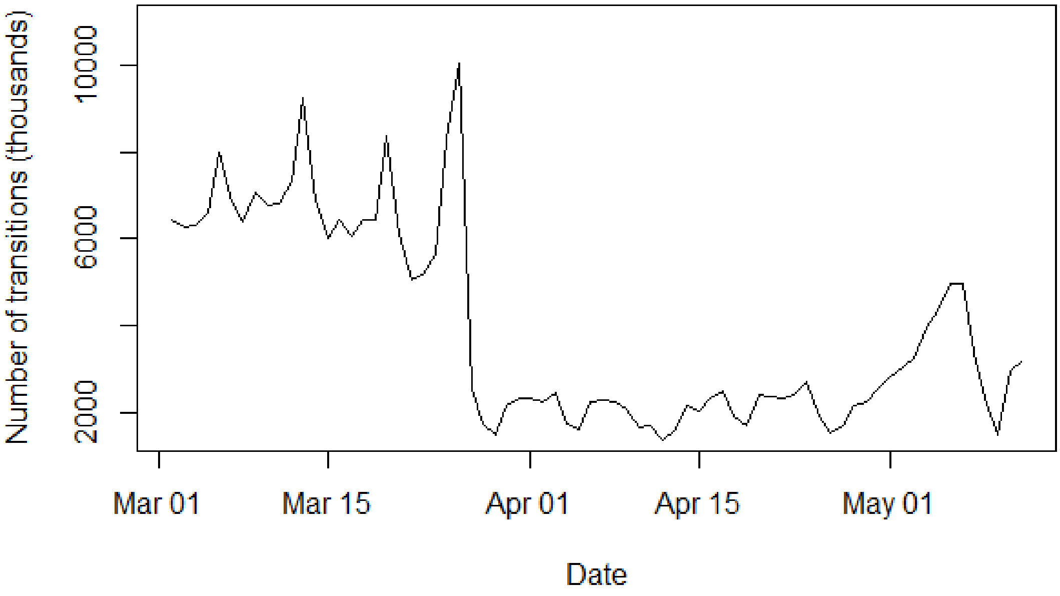

Figure 2. Mobile phone data (2 March 2020 – 12 May 2020)

The population of South Africa is approximately 58 million, and yet the highest total number of inter-district municipality transitions on any given day was approximately 10 million (seen in Figure 2). It should be noted that the same individual can be responsible for multiple transitions and that some individuals could potentially possess multiple cellular devices. Various literature exists on the use of mobile phone data to estimate population numbers (see e.g. [21]).

Despite the quality of available hardware (64GB RAM PC), this process proved highly computationally expensive due to the number of comparisons that need to be run on billions of lines of data in order to create a spatial weight matrix for each day in the time period. The mobile phone data has only been fully converted to administrative level 3 (see Table 1). The rest of the discussion of this data will thus be restricted to this spatial resolution.

2.3 South Africa's lockdown levels

To quell the spread and impact of the COVID-19 pandemic, the South African government instigated one of the strictest lockdowns in the world. This particular lockdown strategy is structured around different "levels" of lockdown, each of which brings different restrictions (with level 5 being the highest and placing restrictions on nearly all forms of travel to all citizens except for those classified as essential workers). The various levels as well as the dates for which they were active are given in Table 2. Note that for this paper we only consider the lockdown until the end of Level 3 due to data availability only over this period.

Table 2. South Africa lockdown levels and dates

Level	Date	Restrictions
Business as usual	1 March 2020 - 26 March 2020	No restrictions
Level 5	27 March 2020 - 30 April 2020	Essential services only otherwise all confined to place of residence. No inter-provincial movement, except for transportation of goods and exceptional circumstances e.g. funerals. Public and private transport restricted to certain times of the day, with limitations on vehicle capacity
Level 4	1 May 2020 - 31 May 2020	More sectors permitted with restrictions, including mining, and partial e-Commerce allowed. Public places (such as religious, cultural, recreational facilities) and the tourism sector remain closed and gatherings prohibited. All confined to place of residence from 8pm-5pm. No local (between metropolitan areas or districts) or inter-provincial movement of people, except for permitted reasons e.g. returning for alert level 4 operations. All borders remain closed except for designated ports of entry for restricted home affairs operations and for the transportation of fuel, cargo and goods. Public and private transport may operate at all times of the day, with limitations on vehicle capacity
Level 3	1 June 2020 - 17 August 2020	More sectors permitted including take away restaurants, e-commerce and delivery services and global business services. Public places and tourism opened and gatherings and sporting activities permitted but all subject to restrictions. All confined to place of residence from 11pm-4pm. No inter-provincial movement of people, except for transportation of goods, exceptional circumstances and other permitted reasons. : Public and private transport may operate at all times of the day, with limitations on vehicle capacity

As non-pharmaceutical interventions (such as the lockdown) are eased the population is allowed to become more mobile. Naturally this will have an impact on the transmission rate of the virus and thus this temporal element must be included in some manner. In this paper we split the data temporally on the date ranges given in Table 2 and set up a spatial weight matrix for each level of lockdown to study how mobility patterns changed.

3 METHODOLOGY

3.1 Literature review

When a particular phenomenon exhibits evidence of spatial dependence, this dependency must be taken into account when modelling to minimise the risk of producing biased results [22, 23]. In the case of an infectious disease that is spread through physical contact and near proximity, it is clear that locations that are situated closer together (or rather the inhabitants of these locations) will play a larger role in determining their respective infection rates than locations that are farther apart. To incorporate this fact, spatial models allow spatial units to be more strongly (or weakly) correlated with one another based on some select criteria that is deemed suitable for the phenomenon being modelled. This is achieved through the use of a spatial weight matrix (sometimes called a "spatial mobility matrix") usually denoted by $\boldsymbol{W}$ [22, 24, 25, 26, 23, 27].

Definition 1 (Spatial weight matrix)**.** Let $S = \{1, 2, \ldots, n\}$ be a set of spatial units. A spatial weight matrix is an $n \times n$ matrix $\boldsymbol{W} = [w_{ij}]$ satisfying [1,4]

1.$w_{ij} \geq 0$

2.$\sum_{j=1}^{n} w_{ij} = 1 \quad \forall \quad i \in S$

This matrix is formally defined as an expression of spatial dependency between spatial units [22, 24, 25, 26]. Simply put, the spatial weight matrix is constructed in such a way so that entry w_{ij} quantifies the amount of spatial influence that spatial unit i exerts on spatial unit j [22, 24, 25, 26].

Such matrices are frequently restricted to being symmetrical to simplify estimation. However, symmetry is not required and can result in a less realistic representation of spatial dependency [22, 24, 25, 26]. Another common convention is that $w_{ii} = 0$ for all i to exclude the possibility of so-called "self-influence" [22, 24, 26]. Non-zero diagonal entries can however be included and are interpreted as quantifying the resistance that each spatial unit has against influence from the other spatial units [25, 26]. Performing row-standardisation on the matrix allows the connectivity of different spatial units to be compared [24, 26].

Spatial weight matrices are most commonly used in the fields of econometrics and spatial statistics [25]. Recently however, they have become popular in the field of spatial epidemiology and have been used to model various diseases including dengue, malaria, foot and mouth disease [28, 29, 30, 31] and most recently COVID-19 [32, 33]. There are relatively few established guidelines with regards to constructing a spatial weight matrix [22, 26, 23, 27], however, the construction of these matrices has seen some improvement as well, with greater emphasis being placed on creating matrices that offer an accurate representation of human mobility. Simpler models rely on measures such as distance, contiguity or adjacency [22, 25, 23, 27, 28, 29, 30, 31, 33] while more complex ones are able to use mobile phone data [33] and geostatistical information [24, 27]. Accurately specifying these matrices is a non-trivial problem, as discussed in [23]. Most recently, Ejigu et al. proposed a methodology through which both distance and covariate information can be utilized [23].

Given the importance of correctly specifying the spatial weight matrix, and the fact that there are often multiple sources of spatial data available on hand, it becomes necessary to develop some means of comparing spatial weight matrices. Specifically, it is necessary to compare the insights that can be derived from different spatial weight matrix definitions. In recent years this comparison has been achieved either through the use of measures of spatial autocorrelation, such as Moran's I [30], or through more specialised methods local to the field of spatial statistics [34, 35]. In this paper, we adapt an idea initially presented by Garrison and Marble [36], whereby principal component analysis is used to reduce the dimensionality of candidate spatial weight matrices. We then use hierarchical clustering to derive a clustering solution for the spatial unit principal scores. This allows for a more in-depth comparison of the information provided by these connectivity matrices, as opposed to simply comparing their visible structure.

3.2 Spatial weight matrices

Selecting an optimal spatial weight matrix is often reliant on the use of a priori information and experience. In this paper the emphasis is on comparing the implications for different spatial weight matrices and the varying types of spatial associations that they represent. We now discuss the spatial weight matrix construction approaches used in this paper.

3.2.1 Method 1: Distance method

The exponential distance definition of a spatial mobility matrix is used frequently in studies involving spatial correlation, and is a popular choice in spatial econometrics [22, 25, 23, 27]. As previously mentioned however, the concepts of distance, contiguity and adjacency do not necessarily offer the most accurate or realistic representation of human mobility. In this paper we include this model in order to draw comparisons between it and more data-driven models. The entries of the spatial weight matrix are given by

$$w_{ij} = \exp(-d_{ij}) \tag{1}$$

where d_{ij} is the Euclidean distance between the centroids of district municipality i and j. Diagonal entries are set to 0 to remove the possibility of so-called "self-influence", and all rows are standardized to sum to 1 to facilitate comparisons between different spatial units. Both of these restrictions were maintained for all matrices in this paper. Under this model, districts are most strongly spatially correlated with the districts that are closest to them geographically.

3.2.2 Method 2: Mobile network method

The mobile network data indicates the number of individuals that travelled from district municipality I to district municipality J on a given day t. These entries are used to construct the spatial weight matrix as follows,

$$w_{ij}^{(t)} = M_{IJ}^{(t)}. \tag{2}$$

This model expresses spatial weights as a function of the amount of flux (both in and out) occurring at a spatial location, and is sometimes referred to as a spatial interaction matrix [26]. District municipalities where more (less) individuals travelled to other district municipalities will thus have a larger (smaller) effect on other district municipalities.

3.2.3 Method 3: Weighted Facebook data method

In order to create a spatial mobility matrix using the Facebook data, we use the same approach of Ejigu et al. [23]. This takes into account proximity as well as covariate information which is spatially dependent. The entries of the the spatial weight matrix are given by

$$w_{ij}^{(t)} = \exp\left(-\left(\alpha \cdot |F_i^{(t)} - F_j^{(t)}| + (1-\alpha) \cdot d_{ij}\right)\right) \tag{3}$$

where $F_i^{(t)}$ is the mobility of district municipality i at time t, scaled by population size (the covariate information), d_{ij} is the Euclidean distance between the centroids of district municipality i and j, and $\alpha \in (0,1)$ is a control parameter indicating the amount of weight that should be given to the covariate term. The control parameter α was set to 0.6 in this paper to allow for the covariate data to play a slightly more prominent role in the estimation process without disregarding the importance of distance. The parameter captures that we are making an assumption that the Facebook data can be used to capture transitions between regions even though it is isolated location data. The value of 0.6 gives the weighted calculation a slight nudge towards the Facebook data. Note that if $\alpha = 0$ then the model simplifies to the exponential distance model.

The Facebook mobility data for each district municipality was scaled using population size in order to account for the fact that increased mobility in a given district is more (less) influential to neighbouring districts if the population size is large (small). This was also done in order to restore some of the variation in the data that was likely lost when the data was censored to a lower spatial resolution.

3.2.4 Method 4: Scaled Facebook data method

A final spatial weight matrix was constructed based on another variation of the exponential distance model. For this matrix, the rows of the exponential distance matrix are scaled using the (unscaled) Facebook mobility data. For example, if the mobility within district municipality i was 20% lower than the baseline, then the entire row i is multiplied by 0.8. Each entry in the exponential distance matrix is thus scaled by some number in (0,2). The entries in the matrix are given by

$$w_{ij}^{(t)} = \left(1 + F_i^{(t)}\right) \cdot \exp(-d_{ij}). \tag{4}$$

This construction allows the exponential distance matrix to be scaled such that the spatial influence of more (less) mobile district municipalities is increased (decreased). This also renders the exponential distance matrix non-symmetric, which should offer a more realistic representation of spatial influence.

3.3 Principal Component Analysis

Principal component analysis (PCA) is a statistical technique that aims to derive a parsimonious representation of a given dataset by deriving an orthogonal linear transformation of the data [37]. In standard PCA, the only hyperparameter that needs to be selected is the number of principal components, which is primarily dependent on the cumulative proportion of variance in the data that the user wishes to retain. For this paper, the number of principal components was chosen such that 75% of the variation in the data was maintained. The full discussion of PCA and its various extensions is left to the existing literature (see e.g. [37]).

3.4 Hierarchical clustering

Hierarchical clustering is an unsupervised machine learning technique that allows the user to group together data points in an attempt to uncover sets of observations that share similar characteristics [37]. This is achieved by procedurally grouping together those observations that are most similar to each other based on some selected measure of dissimilarity, referred to as a "linkage" [37]. The number of retained clusters can then be selected either using some measure of cluster (dis)similarity or a pre-selected value. We use agglomerative clustering, which additionally requires the selection of a method through which the dissimilarity of separate clusters is calculated. A full discussion on hierarchical clustering may be found in

[37].

Herein, we chose the number of clusters to be identical to the number of principal components. Complete linkage was used to calculate the difference between clusters at each iteration. Single and average linkage displayed a propensity for resulting in clusters that were very large. This was most likely due to the fact that single linkage considers the minimum distance between clusters at each iteration, thus regarding clusters as more similar in general. Complete linkage considers the maximum distance between clusters and thus considers clusters to be more distinct. Average linkage is the average of these two extremes.

4 RESULTS

Figure 3 shows the 52 district municipalities of South Africa. The four largest cities in the country are Tshwane, Johannesburg, Durban and Cape Town, situated in the City of Tshwane, City of Johannesburg, eThekwini and City of Cape Town district municipalities respectively as indicated in colour in the figure. These four cities are the focal point of economic activity and travel in the country, and it is thus logical that they would play a substantially larger role in the transmission of the virus than other municipalities.

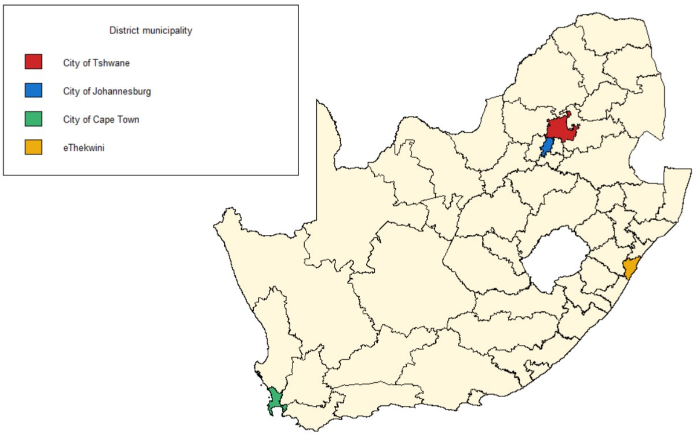

Figure 3. South African district municipalities (locations of four largest cities indicated in colour)

4.1 Method 1: Distance method

Figure 4 shows the weights of the exponential distance weight matrix. Since the entries are calculated based only on the Euclidean distance between the district municipalities (and no additional information), there are no significantly large weights present. As temporal information is not included, this method produces only a single spatial weight matrix.

This spatial weight matrix required the largest number of principal components, namely 14, in order to explain 75% of the variation in the data. This is most likely due to the lack of any form of auxiliary data or information that could be used to better describe the relationship of the district municipalities. The result of hierarchical clustering on the principal component observations is given in Figure 4.

4.2 Method 2: Mobile network method

Due to the great computational cost of the mobile phone data prohibiting analysis at administrative level 2, we discuss the results obtained for administrative level 3. Figure 5 shows the resulting spatial weight matrix for every level of lockdown that the mobile phone data spans. This spatial weight matrix identifies very strong spatial associations over relatively shorter distances (indicated by the yellow lines). These strong correlations appear to cluster around the edges of the country, with locations in the centre of the

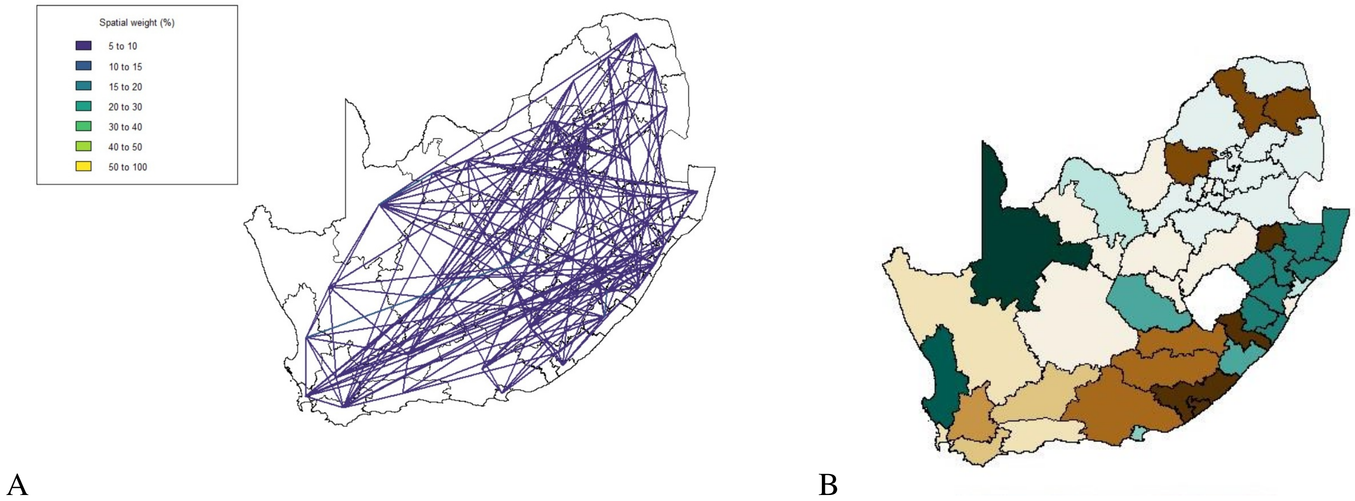

Figure 4. Method 1 A. Spatial weights (weights ≤ 5 not shown), B. Complete linkage clustering (clusters indicated by colours)

country displaying less spatial association overall.

We note that there are strong spatial associations that do not appear to be associated with any of the four major cities in the country. In particular, we note strong associations in the North-Western region of the country as well as some spatial associations across Lesotho (a neighboring country that is landlocked by South Africa, shown in Figure 5D).

A notable drawback of data being at such a high spatial resolution is that it becomes very difficult to cluster locations in a meaningful way. At administrative level 3 there are 213 spatial units to consider. In order to explain just 75% of the variation in this data one requires approximately 70 principal components. Such a high number of clusters does not lend itself to easy interpretation and thus we are reliant on visualizations such as those in Figure 5. In order to facilitate further analysis, the data could be aggregated to a lower spatial resolution.

4.3 Method 3: Weighted Facebook data method

This matrix construction incorporates both the Facebook population mobility data and the population size for each district municipality into the spatial weights for each district municipality pair. Figure 6 shows the resulting matrix for each level of lockdown. By allowing both mobility and population size to play a role in this matrix, the strong spatial association between the four largest cities in South Africa is identified, despite the large geographical distance between them. If only Euclidean distance had been taken into account, this association would have been missed, as with Method 1.

This spatial weight matrix required 9 principal components to explain 75% of the variation in the data. Figure 7 shows the results of applying hierarchical clustering to the principal component observations.

4.4 Method 4: Scaled Facebook data method

This spatial weight matrix was constructed as a potentially more realistic alternative to the exponential distance matrix. Despite containing a temporal element (in the form of daily mobility measurements retrieved from the Facebook data), the results for this matrix do not show any significant change across the various levels of lockdown. Figure 8 shows the elements of the spatial weight matrix.

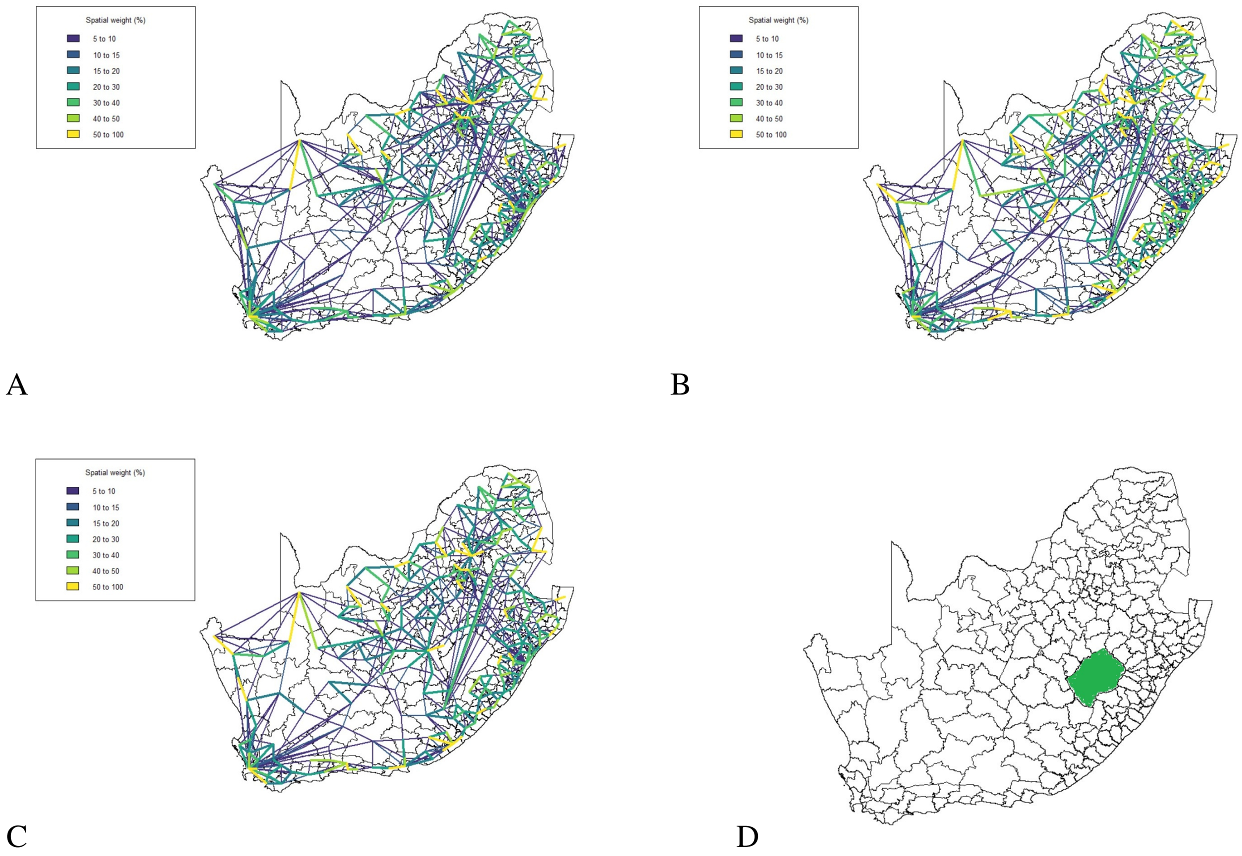

Figure 5. Method 2 spatial weight matrix entries (weights ≤ 5 not shown) A. Business as usual, B. Level 5, C. Level 4, and D. South Africa at Administrative level 3 (neighboring country Lesotho in green)

Clustering performed on this matrix was more successful and intuitive. Only 7 components were required to explain 75% of the variation in the data. Figure 8 shows the clustering solution.

5 DISCUSSION

The results shown in Section 4 illustrate a number of ways to construct spatial weight matrices from mobility data. For the standard exponential distance method, it is clear from Figure 4 that the clustering solution on this spatial weight matrix is not ideal. There are far too many clusters and the clustering solution reveals no clear interpretation. Although the initial matrix construction used only the distances between district municipalities, district municipalities that were located closer together were not generally clustered together.

The entries of the spatial weight matrix constructed using the mobile phone data, shown in Figure 5, reveal strong spatial associations over relatively short distances. The four focal largest cities in the country are clearly identified as hubs for high mobility but there are other regions, particularly those situated on or near the borders of the country, that showcase highly concentrated mobility. A possible explanation for these strong spatial associations being observed far away from cities is the existence of mining activity in these areas. Given that South Africa has a very large and widespread mining sector it seems only reasonable that any model with a spatial element should strive to incorporate these associations.

The four largest cities in South Africa are Tshwane, Johannesburg, Cape Town and Durban, situated

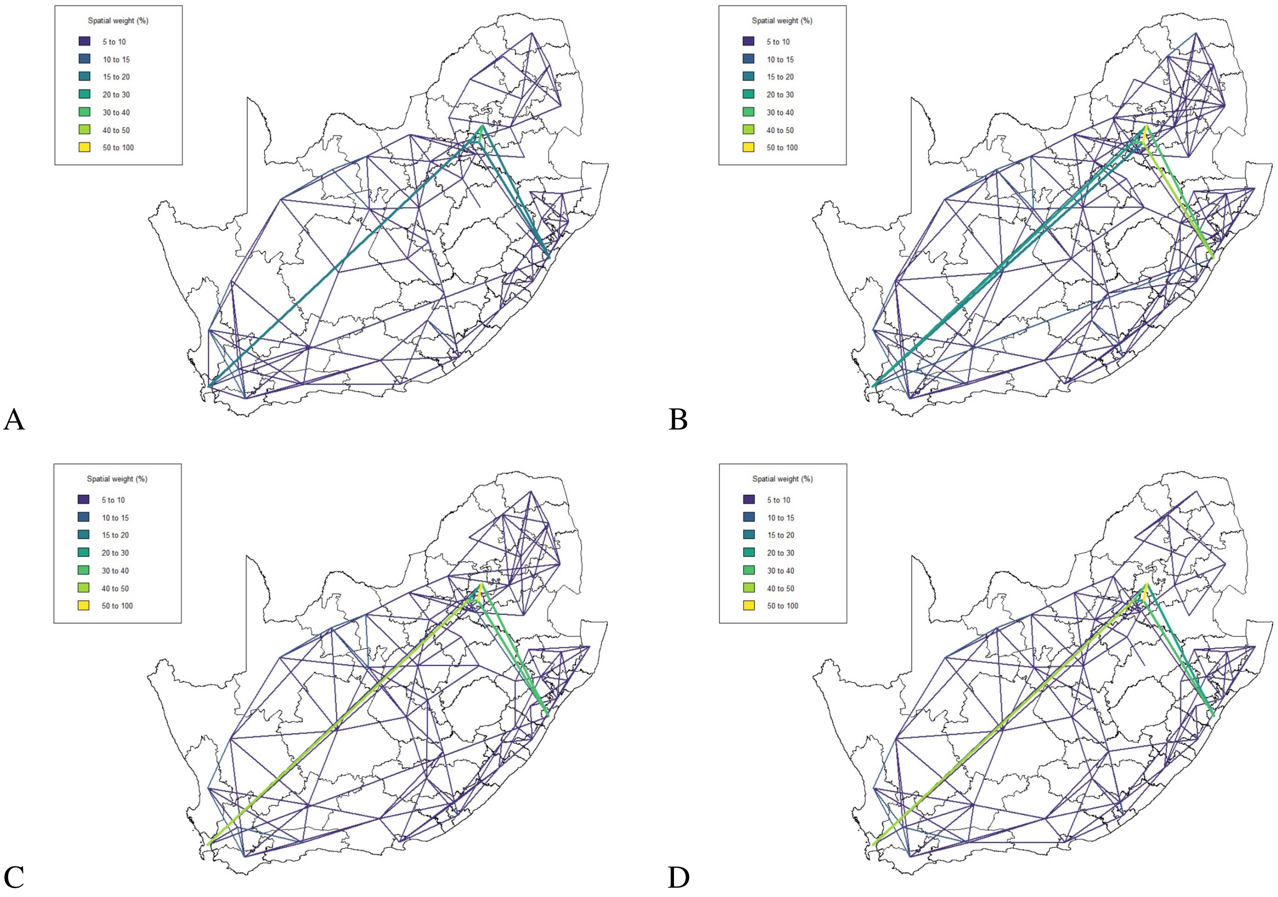

Figure 6. Method 3 spatial weight matrix entries (weights ≤ 5 not shown) A. Business as usual, B. Level 5, C. Level 4, and D. Level 3

in the City of Tshwane, City of Johannesburg, eThekwini and City of Cape Town district municipalities respectively as shown in Figure 3. The results in Figure 6 show a large spatial association between these locations prior to the implementation of level 5 lockdown. Under level 5 restrictions, when the spatial influence of most district municipalities decreased, the spatial influence between these four locations became more pronounced by comparison. This most likely indicates that while smaller district municipalities were less active due to restrictions, these four were comparatively more active and still saw a sizable amount of travel between them. This seems feasible, given that these locations are the focal points for economic activity in the country and thus could not reasonably become "immobile". As restrictions were lifted, these spatial weights were still significantly larger than those for other district municipalities, indicating that, despite restrictions being eased, the spatial influence between these four places is still significantly stronger than before the lockdown. It is also apparent that the spatial influence between less influential district municipalities has not returned to the level that they were during business as usual (pre-lockdown). Figure 7 shows that the district municipalities housing the four largest cities are all either clustered together or in clusters of their own. Other district municipalities are generally clustered together based on their distance to one another. This clustering solution indicates that the four largest cities are significantly different from the locations around them. This spatial weight matrix is thus able to pinpoint the fact that these locations play a potentially larger role in spatially-dependent phenomena such as the spread of a virus.

The clustering results for Method 4 spatial weight matrix, shown in Figure 8, do not display any significant

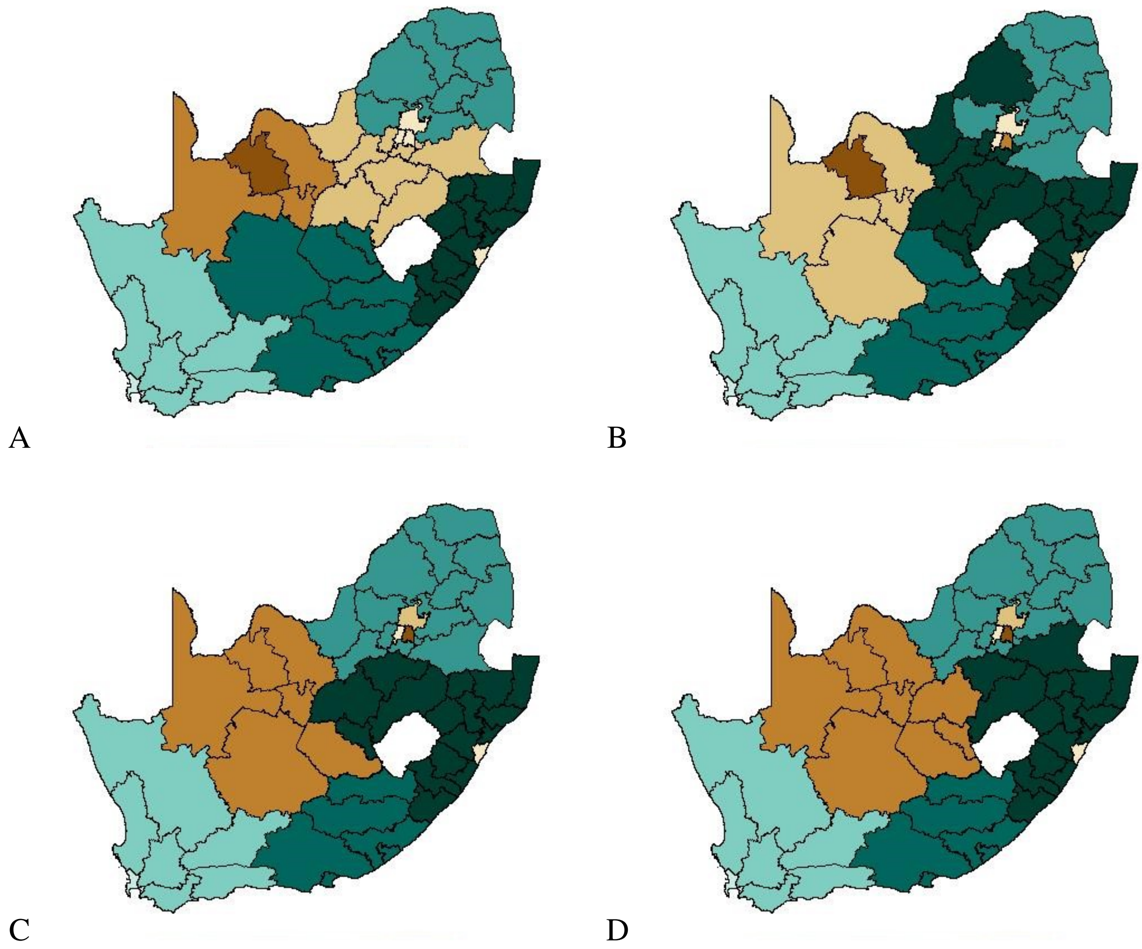

Figure 7. Method 3 complete linkage clustering results A. Business as usual, B. Level 5, C. Level 4, and D. Level 3

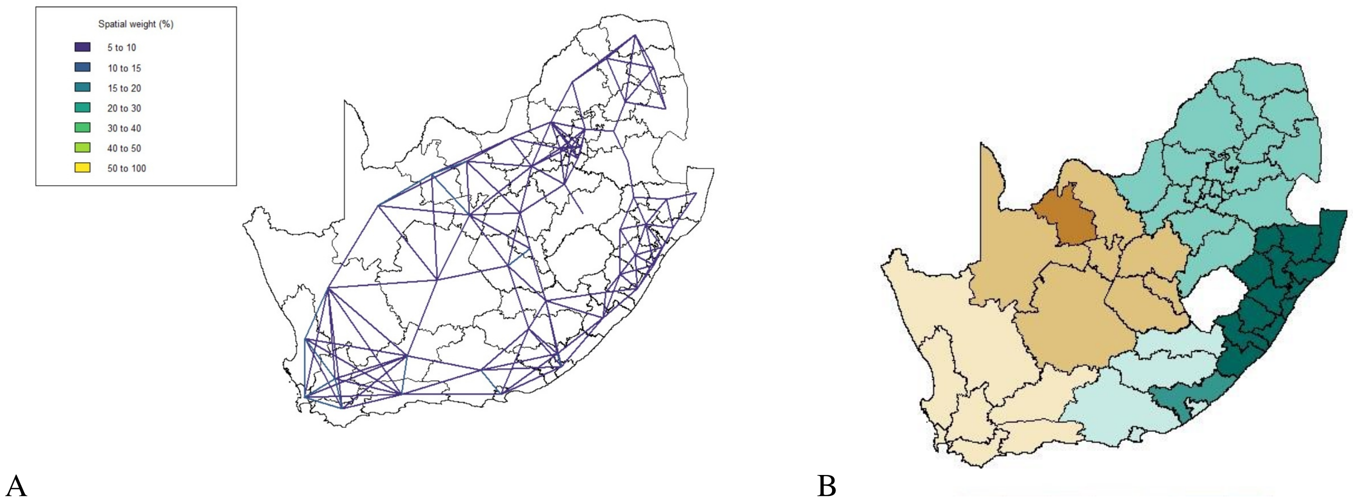

Figure 8. Method 4 A. Spatial weights (weights ≤ 5 not shown), B. Complete linkage clustering (clusters indicated by colours)

changes over the various levels of lockdown. Figure 8 also shows that the clusters that are formed for this spatial weight matrix are clearly based primarily on distance but with the auxiliary Facebook data aiding in constructing more finite and sensible clusters. Interestingly, we notice a district municipality that has been classified into a cluster on its own. When inspecting the results for the other spatial weight matrices we note that this district municipality has previously also been identified as its own cluster and was shown to have strong spatial associations for Method 2. Upon further inspection we note this district municipality houses several mines. Similar to with Method 2, this spatial weight matrix is able to identify location associations that go unnoticed when relying on simple concepts such as Euclidean distance.

Table 3 provides a summary of the methods used in this paper, their strengths and weaknesses, and their usability. This paper shows that there are different ways to represent spatial data. These can offer a variety of insights and capture different relationships. For example, the spatial weight matrix created using Method 3 data emphasizes the prominent role of focal points in population activity. However, the spatial weight matrix constructed using Method 4 offers a scaled and smoothed way to use distance to indicate which locations have a higher spatial influence on one another. These two spatial weight matrices use the same spatial data (i.e. the Facebook for good data), but offer vastly different interpretations of spatial influence. Finally, the interpretations that were able to be made from the mobile phone data indicates that there are many potentially strong spatial associations at shorter distances that can only be identified when inspecting data at a high spatial resolution.

Each of these representations can be seen as valid and are complementary with regards to the insight they offer. Depending on the specific phenomenon under study, an argument could be made for any of them. In the case of a pandemic such as COVID-19, which affects not only congregated communities but allows for consequences to be felt across an entire country, an argument can easily be made for every single representations to be incorporated in some fashion.

Table 3. Spatial weight matrices comparison

Spatial weight matrix	*Pro*	*Con*	*Interpretation/Contribution*
Method 1 - Distance	Simple to construct and understand Used often in literature	Less realistic Inadequate for clustering Lacks temporal element	Convenient to use and easy to understand and interpret. Not realistic enough for real insight.
Method 2 - Mobile network	High spatial resolution Large amounts generated passively by mobile device users	Computationally expensive Difficult to obtain Not representative Privacy concerns	Captures strong spatial associations over relatively short distances. Allows for the identification of patterns potentially missed by other methods.
Method 3 - Weighted Facebook data	Freely available data Potentially more representative	Low spatial resolution Lacks specificity	Captures association between focal points of human activity regardless of distance.
Method 4 - Scaled Facebook data	Simple to construct and understand Freely available data Potentially more representative	Lacks temporal elements Low spatial resolution	Adds additional information to previously simplistic model. Additional information improves clustering.

Understanding mobility during the current pandemic is essential. Both the reduction in mobility as well as retained mobility need to be well understood, and depend on reliable data collection. As shown here, data are collected in different ways and are also made available in a variety of formats. Mobility is distributionally different across strata of a region's demographics, with more mobile locations likely to result in higher disease transmission. Higher resolution mobility data is important to capture these differences in more detail. Even so, the spatial resolution at district municipality captures these nuances of the movement under each lockdown level, and shows that significant movement still took place due to the vulnerability of a large portion of South Africa's population.

The possibility of micro-spatial estimation (small area estimation) is something to further investigate. Making use of demographic covariates and transport networks, as well as mobile network coverage maps could provide connectivity matrices at higher spatial resolution, ideally at ward level. This allows for micro-scale modelling of COVID-19 spread. This approach will also allow for privacy while increasing spatial resolution and providing deeper coverage in a region.

Google mobility data is also available[5] but only at provincial level (Administration level 1) for South Africa. This spatial resolution is too sparse to consider estimation down to ward level, especially if mobility data is available at administrative level 2. However, combining mobility data of different spatial resolutions could also be considered taking advantage of the strengths of each dataset.

The computational aspects of dealing with mobility data should not be overlooked. Spatial weight matrices can become very large, depending on the number of spatial regions under consideration. Herein the matrices were not sparse, thus not allowing for sparse representation.

6 CONCLUSION

COVID-19 spreads spatially and thus the importance of mobility data for COVID-19 modeling should not be disregarded. Ideally, the raw data from the mobile network providers and Facebook, if available, could provide individual movements, allowing for accurate construction of spatial weight matrices. This data could be anonymised and shared. The use of movement data in epidemiology is becoming an important covariate to include, without which the spread can only be modelled in isolated regions. The social nature of humans is unavoidable. Simple spatial weight matrix construction techniques, such as only taking into account distances, are not always ideal when the spatial associations being captured is dependent on covariates which are not only proximity based. This is made clear by what a poor job Method 1 does when used as the basis of clustering. The methods presented herein and the results shown also enable epidemiological modellers in considering how to incorporate spatial relationships in models. This is seldom done due to limited mobility information as well as modelling complexities it introduces. However, the improved accuracy in model outcomes will ultimately balance out computational complexities. The paper provides insights into mobility data availability, representability as well as construction for use in spatial modelling. Future research should investigate estimation to a higher spatial resolution as well as the effect of spatial resolution in spatial epidemiological modelling.

ACKNOWLEDGEMENTS

We thank Gerbrand Mans from the CSIR for data aggregation assistance as well a Bruce Medallo from WITS for mobility data advice.

The financial assistance of the National Research Foundation (NRF) towards this research is hereby

[5] See the following link

acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

This research is also funded by Canada's International Development Research Centre (IDRC) (Grant No. 109559-001).

The use of the Centre for High Performance Computing (`www.chpc.ac.za`) also made this work possible.

DATA AVAILABILITY STATEMENT

The mobile network data used in this study is not directly available without approval, so cannot be shared directly with the paper. Thank you to the NICD, South Africa for providing access this data for the COVID-19 response in South Africa.

AUTHOR CONTRIBUTION

Conceptualisation: All; Data Curation: AP, IFR, ZK, PD; Formal Analysis: AP, IFR; Funding Acquisition: PD; Investigation: AP, IFR; Methodology: AP, IFR; Project Administration: ZK, PD; Writing – original draft: AP, IFR; Writing – review editing: All.

CONFLICTS OF INTEREST

Author Sibu Makhanya is employed by IBM, South Africa. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

[1] Huang X, Li Z, Jiang Y, Ye X, Deng C, Zhang J, et al. The characteristics of multi-source mobility datasets and how they reveal the luxury nature of social distancing in the us during the covid-19 pandemic. *International Journal of Digital Earth* **14** (2021) 424–442.

[2] Finger F, Genolet T, Mari L, de Magny GC, Manga NM, Rinaldo A, et al. Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks. *Proceedings of the National Academy of Sciences* **113** (2016) 6421–6426.

[3] Bengtsson L, Gaudart J, Lu X, Moore S, Wetter E, Sallah K, et al. Using mobile phone data to predict the spatial spread of cholera. *Scientific Reports* **5** (2015) 1–5.

[4] Cummings DA, Irizarry RA, Huang NE, Endy TP, Nisalak A, Ungchusak K, et al. Travelling waves in the occurrence of dengue haemorrhagic fever in Thailand. *Nature* **427** (2004) 344–347.

[5] Wesolowski A, Qureshi T, Boni MF, Sundsøy PR, Johansson MA, Rasheed SB, et al. Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proceedings of the National Academy of Sciences* **112** (2015) 11887–11892.

[6] Ruktanonchai NW, DeLeenheer P, Tatem AJ, Alegana VA, Caughlin TT, zu Erbach-Schoenberg E, et al. Identifying malaria transmission foci for elimination using human mobility data. *PLoS Computational Biology* **12** (2016) e1004846.

[7] Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, et al. Quantifying the impact of human mobility on malaria. *Science* **338** (2012) 267–270.

[8] Asgari F, Gauthier V, Becker M. A survey on human mobility and its applications. *arXiv preprint arXiv:1307.0814* (2013).

[9] Zhou Y, Lau BPL, Yuen C, Tunçer B, Wilhelm E. Understanding urban human mobility through crowdsensed data. *IEEE Communications Magazine* **56** (2018) 52–59.

[10] Toch E, Lerner B, Ben-Zion E, Ben-Gal I. Analyzing large-scale human mobility data: a survey of machine learning methods and applications. *Knowledge and Information Systems* **58** (2019) 501–523.

[11] Ballas D, Clarke G, Dorling D, Eyre H, Thomas B, Rossiter D. Simbritain: a spatial microsimulation approach to population dynamics. *Population, Space and Place* **11** (2005) 13–34.

[12] Pfeffermann D, et al. New important developments in small area estimation. *Statistical Science* **28** (2013) 40–68.

[13] Ekong I, Chukwu E, Chukwu M. Covid-19 mobile positioning data contact tracing and patient privacy regulations: exploratory search of global response strategies and the use of digital tools in Nigeria. *JMIR mHealth and uHealth* **8** (2020) e19139.

[14] Varsavsky T, Graham MS, Canas LS, Ganesh S, Pujol JC, Sudre CH, et al. Detecting COVID-19 infection hotspots in England using large-scale self-reported data from a mobile application: a prospective, observational study. *The Lancet Public Health* **6** (2021) e21–e29.

[15] Oliver N, Lepri B, Sterly H, Lambiotte R, Deletaille S, De Nadai M, et al. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle (2020).

[16] Peixoto PS, Marcondes D, Peixoto C, Oliva SM. Modeling future spread of infections via mobile geolocation data and population dynamics. an application to COVID-19 in Brazil. *PloS one* **15** (2020) e0235732.

[17] Gao S, Rao J, Kang Y, Liang Y, Kruse J, Dopfer D, et al. Association of mobile phone location data indications of travel and stay-at-home mandates with COVID-19 infection rates in the US. *JAMA network open* **3** (2020) e2020485–e2020485.

[18] Zhou Y, Xu R, Hu D, Yue Y, Li Q, Xia J. Effects of human mobility restrictions on the spread of covid-19 in Shenzhen, China: a modelling study using mobile phone data. *The Lancet Digital Health* **2** (2020) e417–e424.

[19] Xiong C, Hu S, Yang M, Luo W, Zhang L. Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *Proceedings of the National Academy of Sciences* **117** (2020) 27087–27089.

[20] Grantz KH, Meredith HR, Cummings DA, Metcalf CJE, Grenfell BT, Giles JR, et al. The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nature communications* **11** (2020) 1–8.

[21] Sakarovitch B, Bellefon MPd, Givord P, Vanhoof M. Estimating the residential population from mobile phone data, an initial exploration. *Economie et Statistique* **505** (2018) 109–132.

[22] Stakhovych S, Bijmolt TH. Specification of spatial models: A simulation study on weights matrices. *Papers in Regional Science* **88** (2009) 389–408.

[23] Ejigu BA, Wencheko E. Introducing covariate dependent weighting matrices in fitting autoregressive models and measuring spatio-environmental autocorrelation. *Spatial Statistics* **38** (2020) 100454.

[24] Getis A, Aldstadt J. Constructing the spatial weights matrix using a local statistic. *Geographical Analysis* **36** (2004) 90–104.

[25] Anselin L. *Spatial econometrics: methods and models*, vol. 4 (Springer Science & Business Media) (2013).

[26] Bavaud F. Models for spatial weights: a systematic look. *Geographical Analysis* **30** (1998) 153–171.

[27] Aldstadt J, Getis A. Using AMOEBA to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis* **38** (2006) 327–343.

[28] Malik R, Deardon R, Kwong GP. Parameterizing spatial models of infectious disease transmission that incorporate infection time uncertainty using sampling-based likelihood approximations. *PloS One* **11** (2016) e0146253.

[29] Brown GD, Oleson JJ, Porter AT. An empirically adjusted approach to reproductive number estimation for stochastic compartmental models: A case study of two ebola outbreaks. *Biometrics* **72** (2016) 335–343.

[30] Suryowati K, Bekti R, Faradila A. A comparison of weights matrices on computation of dengue spatial autocorrelation. *IOP Conference Series: Materials Science and Engineering* (2018), vol. 335, 012052.

[31] Brown GD, Porter AT, Oleson JJ, Hinman JA. Approximate Bayesian computation for spatial SEIR(S) epidemic models. *Spatial and Spatio-Temporal Epidemiology* **24** (2018) 27–37.

[32] Tagliazucchi E, Balenzuela P, Travizano M, Mindlin G, Mininni PD. Lessons from being challenged by COVID-19. *Chaos, Solitons & Fractals* **137** (2020) 109923.

[33] Huang R, Liu M, Ding Y. Spatial-temporal distribution of covid-19 in China and its prediction: A data-driven modeling analysis. *The Journal of Infection in Developing Countries* **14** (2020) 246–253.

[34] Gao Y, Li T, Wang S, Jeong MH, Soltani K. A multidimensional spatial scan statistics approach to movement pattern comparison. *International Journal of Geographical Information Science* **32** (2018) 1304–1325.

[35] Jin C, Nara A, Yang JA, Tsou MH. Similarity measurement on human mobility data with spatially weighted structural similarity index (SpSSIM). *Transactions in GIS* **24** (2020) 104–122.

[36] Garrison WL, Marble DF. Factor-analytic study of the connectivity of a transportation network. *Papers of the Regional Science Association* (Springer) (1964), vol. 12, 231–238.

[37] Friedman J, Hastie T, Tibshirani R, et al. *The Elements of Statistical Learning*, vol. 1 (Springer series in statistics New York) (2001).

APPENDIX

Facebook for good data calculation

Let u represent a single individual and $U_{t,i}$ represent district municipality i at time t. The total number of Bing tiles visited by inhabitants of district municipality i is then

$$total\_tiles(U_{t,i}) = \sum_{u \in U_{t,i}} min\left(tiles(u), 200\right).$$

Note that the maximum number of Bing tiles visited that a single individual can contribute is restricted to 200. In order to preserve user privacy, an error term was included by drawing from a Laplace distribution with parameters 0 and $\frac{F}{\epsilon}$ where $F =$ sensitivity parameter and $\epsilon =$ noise parameter as follows

$$total\_tiles'(U_{t,i}) = total\_tiles(U_{t,i}) + Laplace\left(0, \frac{F}{\epsilon}\right).$$

The average number of tiles per district municipality was then calculated as

$$avg\_tiles'(U_{t,i}) = \frac{total\_tiles'(U_{t,i})}{|U_{t,i}|}.$$

The mobility value for each district municipality and for each day was then finally expressed with respect to the baseline as

$$F_i^{(t)} = \frac{avg\_tiles(U_{t,i}) - baseline\_avg\_tiles'(i, day\_of\_the\_week(t))}{baseline\_avg\_tiles'(i, day\_of\_the\_week(t))}.$$

For further details regarding this data see the following link.