Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-022-01256-8 OPEN

Exploring methods for mapping seasonal population changes using mobile phone data

Data accurately representing the population distribution at the subnational level within countries is critical to policy and decision makers for many applications. Call data records (CDRs) have shown great promise for this, providing much higher temporal and spatial resolutions compared to traditional data sources. For CDRs to be integrated with other data and in order to effectively inform and support policy and decision making, mobile phone user must be distributed from the cell tower level into administrative units. This can be done in different ways and it is often not considered which method produces the best representation of the underlying population distribution. Using anonymised CDRs in Namibia between 2011 and 2013, four distribution methods were assessed at multiple administrative unit levels. Estimates of user density per administrative unit were ranked for each method and compared against the corresponding census-derived population densities, using Kendall's tau-b rank tests. Seasonal and trend decomposition using Loess (STL) and multivariate clustering was subsequently used to identify patterns of seasonal user variation and investigate how different distribution methods can impact these. Results show that the accuracy of the results of each distribution method is influenced by the considered administrative unit level. While marginal differences between methods are displayed at "coarser" level 1, the use of mobile phone tower ranges provided the most accurate results for Namibia at finer levels 2 and 3. The use of STL is helpful to recognise the impact of the underlying distribution methods on further analysis, with the degree of consensus between methods decreasing as spatial scale increases. Multivariate clustering delivers valuable insights into which units share a similar seasonal user behaviour. The higher the number of prescribed clusters, the more the results obtained using different distribution methods differ. However, two major seasonal patterns were identified across all distribution methods, levels and most cluster numbers: (a) units with a 15% user decrease in August and (b) units with a 20–30% user increase in December. Both patterns are likely to be partially linked to school holidays and people going on vacation and/or visiting relatives and friends. This study highlights the need and importance of investigating CDRs in detail before conducting subsequent analysis like seasonal and trend decomposition. In particular, CDRs need to be investigated both in terms of their area and population coverage, as well as in relation to the appropriate distribution method to use based on the spatial scale of the specific application. The use of inappropriate methods can change observed seasonal patterns and impact the derived conclusions.

A full list of authors and their affiliations appears at the end of the paper.

Introduction

In an increasingly connected world, the need for accurate contemporary information at high temporal resolutions has become essential. Intra-annual information on the spatial distribution of populations is crucial for a multitude of applications, including policy planning, resource allocation, transport logistics, disease control and disaster management among many others (Steenbruggen et al., 2015; Gaughan et al., 2016; Kraemer et al., 2019; Ruktanonchai et al., 2020; Lai et al., 2020b). Historically, such data has been collected as part of smaller interviews or surveys, which are limited by their spatial coverage and frequency of collection (Jiang et al., 2017), whilst potentially biased by recall errors (Palmer et al., 2013).

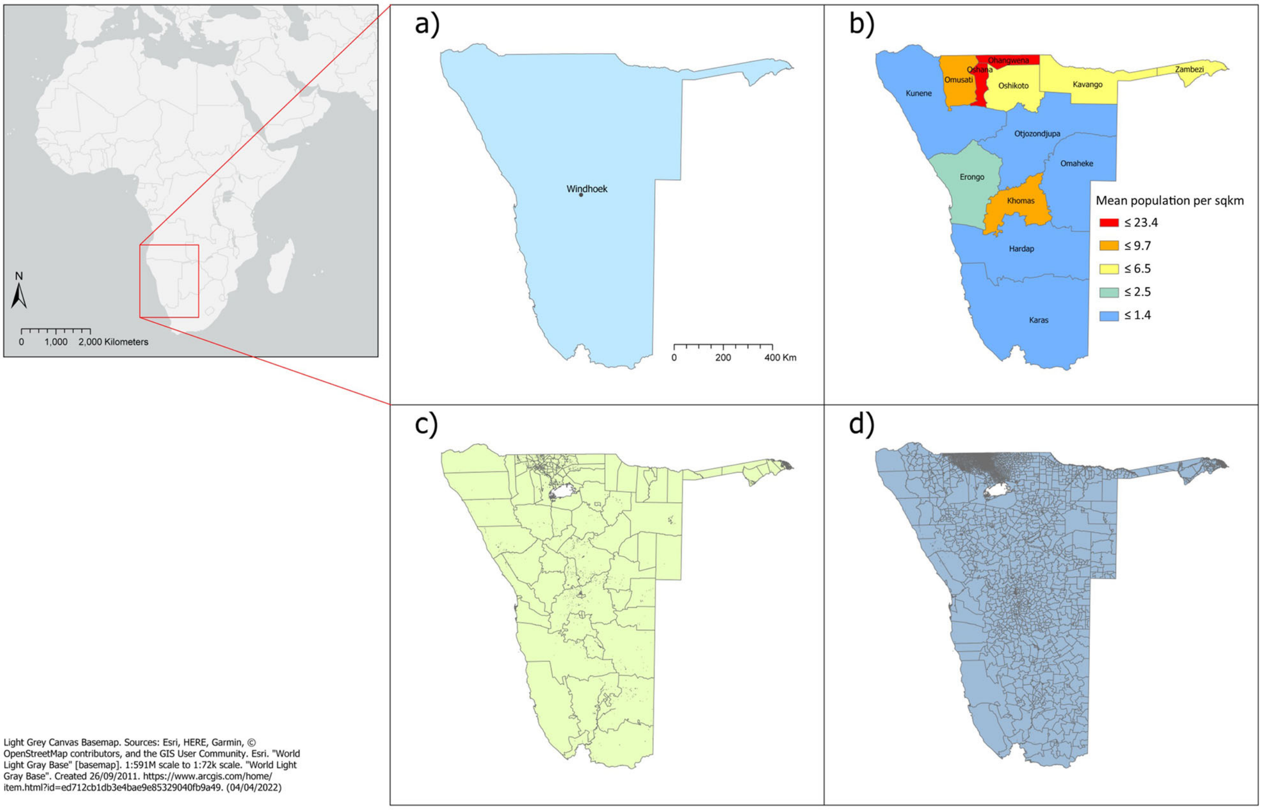
A full census overcomes the limited spatial coverage of smaller surveys and provides essential information about populations. However, census data has its own inherent limitations when used for contemporary planning or assessing short-term human mobility. The typical decadal temporal resolution introduces a significant disadvantage, as the data are temporally fixed and so becomes outdated and less representative of a given population over time. In some countries, census collection can be more irregular and suffer from an even lower temporal resolution which enhances this problem. Beyond this, census data are unable to appropriately assess human population dynamics throughout the year or identify seasonal migration patterns (Davis et al., 2013), which may be needed to provide the most appropriate support to policy and decision makers.

In order to overcome some limitations of the census data, novel sources of human mobility data have been utilised in recent years, such as Global Positioning Systems (GPS) data, to monitor people's movements through space (Vazquez-Prokopec et al., 2013; Tang et al., 2015; Siła-Nowicka et al., 2016). However, the vast growth in ownership of mobile phones has come to the forefront of modern mobility and migration research (Lai et al., 2019b), most prominently in the form of call data records (CDRs) which can cover much greater numbers of people than localised GPS data. CDRs are recorded by carriers for billing purposes and provide the timestamp and position of the base station being used when subscribers make a call, send a text or access the internet (Kanasugi et al., 2013; Wesolowski et al., 2013; Lai et al., 2019b), which then associates the user with a certain tower location.

The ability of CDRs to accurately represent a country's population is therefore of interest to both researchers, decision and policy makers, as it holds the potential to provide human mobility information at higher spatial and temporal resolutions than other data sources (Lenormand et al., 2015). This may allow for events to be responded to more effectively, policies to be based on up-to-date information, and distribution of resources to be streamlined and made more efficient. CDRs have become increasingly adopted to model population distribution (Deville et al., 2014; zu Erbach-Schoenberg et al., 2016) and in mobility research (Xu et al., 2017; Zufiria et al., 2018, Xiao et al., 2019; Zhang et al., 2019; Lai et al., 2020a). Researchers have also used CDRs to assess human mobility relationships with known infectious disease transmission rates, understand connectivity between regions, and inform elimination strategies (Frias-Martinez et al., 2011; Tatem et al., 2014; Ruktanonchai et al., 2016), which further highlights their potential value in complementing more traditional data (Lai et al., 2019a). Some of the shortcomings of CDRs include the reluctance of many mobile network providers to share their data. It can be a lengthy process to draw up precise data agreements therefore most studies work with data from just one provider. Furthermore, CDRs are generally biased due to inhomogeneous market shares of providers (Salat et al., 2020) and their sampling process which is by nature spatially and temporarily non uniform and user dependent (Zufiria et al., 2018). To protect users' identities CDRs typically require anonymising and aggregating. Aggregation is also needed to map the mobile network to the specific units of interest. To be useful for end users, such as decision and policy makers, CDRs must often be related to statistical or administrative units (Fernández and Wu, 2018; Midzi et al., 2018; Poel et al., 2018; Gwitira et al., 2019). Choosing the most appropriate method for distributing the mobile users from the cell tower level into administrative units is therefore the necessary first step in most geographical analysis relying on the use of CDR data. Indeed, it is likely that the chosen distribution method impacts on how accurately the distribution of the underlying population is represented; therefore although particular consideration should be given to this step, in the existing literature this is either overlooked or perhaps the decision is made unconsciously due to lack of knowledge of alternative methods. Whilst easy to implement distribution methods like Point to Polygon allocation or Voronois tessellation are commonly used, a rising number of more complex methods based around maximum likelihood estimation (Ricciato et al., 2017), probabilistic models (Ricciato and Coluccia, Preprint) and signal propagation methods (Koebe, 2020) are being increasingly explored based on the rationale that more advanced and sophisticated distribution methods lead to more accurate results. Koebe (2020) compared Point to Polygon allocation and Voronois tessellation with an augmented Voronois method where the number of settlements within each Voronois were used as weights. They also examined two different signal propagation methods to estimate which towers are most likely to service the different settlements. All five distribution methods were compared both in a simulated country and in Senegal with real-world data from 2013 which were used to estimate unemployment rates. Koebe (2020) got mixed results with no single method consistently outperforming the others. He concluded that while the addition of auxiliary settlement data to inform the Voronois-based distribution improved the model fit, it did not translate into significant efficiency gains for the prediction of unemployment rates. Equally, the value added from using signal propagation methods was neglectable.

The question of what impact the chosen distribution method has on subsequent analysis remains open. In this study we explore the role of the spatial scale of the analysis by using CDR data for Namibia and four methods to distribute user counts from the cell tower level into three administrative unit levels: regions, constituencies and enumeration areas. For each administrative unit level, we compare the distributed users obtained using the different distribution method with the corresponding observed population from the census to assess the performance of distribution methods. While user densities of each unit will never match the true population density, they can be a useful proxy, especially if a strong correlation to a known baseline can be demonstrated. This would also support the point that such data may be used to further analyse the seasonal variation of population distribution. As a potential use case of the distributed user counts, we finally analyse the two methods that match users best/worse for seasonal patterns and cluster them into groups of similar seasonal behaviour. If the observed seasonal patterns and clusters vary, we can conclude that the distribution methods impact outcomes.

Seasonal mobility derived from clustering of CDRs was investigated extensively by Zufiria et al. (2018) on the example of Senegal. They suggest an approach which uses individual trajectory matrices to represent human mobility at the individual (anonymised) level over one year. This allows for the detection of mobility patterns through clustering of individual matrices at the desired spatial and temporal scale. In this study, on the other



Light Grey Canvas Basemap. Sources: Esri, HERE, Garmin, © OpenStreetMap contributors, and the GIS User Community. Esri. "World Light Gray Base" [basemap]. 1:591M scale to 1:72k scale. "World Light Gray Base". Created 26/09/2011. https://www.arcgis.com/home/item.html?id=ed712cb1db3e4bae9e85329040fb9a49. (04/04/2022)

Fig. 1 Overview of Namibia and its divisions. a Administrative level 0 (country) with capital, **b** administrative level 1 (regions) with population density derived from WorldPop (2018), **c** administrative level 2 (constituencies), **d** administrative level 3 (enumeration areas) (GADM, 2022; WordPop, 2018).

Table 1 Characteristics of administrative levels in Namibia.

Administrative unit level	1 (regions)	2 (constituencies)	3 (enumeration areas)
Number of units	13	107	5473
Average unit size (km^2)	63,062	7661	153

hand, we show that seasonal patterns can still be derived even when information on individuals' location is not available. We explore the use of seasonal and trend decomposition using Loess (STL) on CDR data aggregated to administrative units and months in combination with clustering techniques, which to the best of our knowledge has not been done before.

Study area

Namibia is situated in Southern Africa with a population of just above 2 million according to the 2011 census (Namibia Statistics Agency, 2013). As Fig. 1 shows, Namibia is divided into 13 regions (level 1 divisions), 107 constituencies (level 2 divisions) and 5473 enumeration areas (level 3 divisions). The average unit size per level is shown in Table 1. The country has vast amounts of arid land, resulting in clustering of its settlements and population (Linard et al., 2012). Population density therefore varies considerably; besides the capital Windhoek located in the Khomas region in the centre of the country where 16% of the population are located (Pendleton et al., 2014), the majority of people live in the five northern regions of Omusati, Oshana, Oshangwena, Oshikoto and Kavango (see also Fig. 1b). Urbanisation is an ongoing rapid process in Namibia, however, during the time frame we are examining in this study the majority of the population, namely 57%, was still living in rural areas (Namibia Statistics Agency, 2015). The only known seasonal population movements we are aware of were observed in a study about the impact of seasonally varying population numbers on disease incidence estimates by zu Erbach-Schoenberg et al. (2016) and are based on the same CDR data as are used in our study. They found a significant number of people travelling from the capital to the relatively highly populated North of the country in December and returning in January. Similar behaviour, despite at a smaller magnitude, was observed during Easter.

Ownership of mobile phones in households has risen dramatically to 89% in 2013 and is higher in urban households (95%) than in rural households (81%) according to the Namibia Demographic and Health Survey (The Namibia Ministry of Health and Social Services and ICF International, 2014). The two main mobile network providers are Mobile Telecommunications Limited (MTC) and the state owned Telecom Namibia Limited.

CDR and mobile phone tower data

CDR data was obtained from Mobile Telecommunications Limited (MTC), the leading mobile network provider in Namibia with a market share of 76% for 2010–2012. Figure 2 shows cell tower density at level 1 which mirrors the highly populated areas in the capital and the North of the country. In order to comply with the proprietor's terms and conditions, details on tower

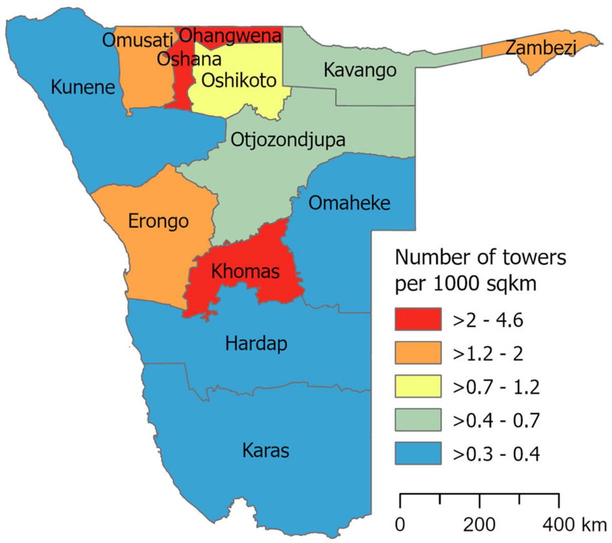


Fig. 2 Cell tower density in Namibia at administrative level 1 (regions; GADM, 2022).

numbers and precise locations cannot be shared. However, since the spatial scale plays a vital role in this study, the following gives an indication of the relative difference of cell tower ratio per administrative level: consider the average number of cell towers per administrative unit at level 1 to be 100%, this percentage drops substantially to 12.5% at level 2 and to only 0.2% at level 3.

The CDR data spans from October 2010 to April 2014 with a total of 72 billion communications (zu Erbach-Schoenberg et al., 2016), each containing an anonymised user ID, the date and time of communication, and the cell tower ID through which the communication was recorded. Mobile users are defined as SIM cards, i.e. each unique anonymised identifier code in the considered CDRs is defined as a user. MTC also supplied approximated circular ranges of cell tower coverage, but no exact information was given on antenna types, frequencies or output power of base transmitter stations. The processing of the raw CDR data was performed by and is described in detail in zu Erbach-Schoenberg et al. (2016); following common practices in literature, the home location of users was assigned to specific cellular towers each day based on their night time location resulting in the number of unique users per day per tower. In the case of a user having no communications for a day, their last/next known night time location was used instead. While the use of night time location is a widely used home detection algorithm, recent research has shown that it is very sensitive to parameter choice, like the exact hours defining night time or the length of observation (Vanhoof et al., 2018). Zu Erbach-Schoenberg et al. (2016) defined night time to the hours between 8 p.m. and 6 a.m. and finally derived daily and monthly user counts per tower which are used in this study. Since CDR data are known to be noisy, monthly counts provide the benefit over daily counts of being sufficiently smoothed. While we will not be able to analyse user movements on a finer temporal level, for example, weekly movements linked to the working week or short-term events like festivals, we believe that months are a useful temporal scale for many policy purposes, for example planning for resource distribution or surveillance and control of infectious diseases.

Within this study, we used daily average user counts per month as a proxy for monthly population distribution. For the temporal analysis we wanted as complete a time series as possible to minimise the risk of noise or outliers, however, due to incomplete data in both 2010 and 2014 we decided to exclude these years in order to eliminate data which may impact on subsequent quality of analysis. This decision left us with three years' worth of data (2011, 2012, 2013) which fulfils the minimum criteria of three complete temporal cycles to perform STL.

We calculated the number of days per month that each tower was active to determine the tower's daily average user count per month. Some towers show clear periods of inactivity, possibly due to towers undergoing repair works. We used the same approach as Vanhoof et al. (2018) and Salat et al. (2020), who came across a similar issue with CDR datasets in France and Senegal and considered these towers as artefacts, therefore they were removed from the analysis. In the absence of common practices we regarded towers with an arbitrary activity level of less than 3 weeks per month (activity below 77%) as unrepresentative of the month as they could potentially miss significant events which may drive population movement. Such towers were hence excluded from that year. Towers in rural areas are generally harder to access for repair works and therefore likely to be inactive for longer periods of time. Providing their location is overlapped by an adjacent tower range, their counts were reassigned to the adjacent tower. If no overlap existed, the counts were dropped to avoid misleading user count fluctuations driven by tower inactivity.

In this manner a 'stable' set of towers was created for 2011, 2012 and 2013 individually, with the number of towers in each month remaining identical throughout each year. Annual tower sets were considered separately because more towers became active over time as a result of expansion of the MTC network. When comparing the number of towers in our annual stable sets specifically, there is a 3% increase between 2011 and 2012, followed by a 1% increase between 2012 and 2013. If just one 'stable' set were to be created for the whole time-series, this would constrain the data to towers which have existed throughout the three study years and ignore data from newly constructed towers in later years. In addition, general growth in mobile phone adoption over the 3-year period is likely to lead to rising user counts, which the detrending algorithm of STL controls for (see section "Analysis and statistical tests", "Decomposition analysis").

Population data

Three different population datasets, gathered from different sources, were used for better estimating the MTC network coverage in terms of population, and thus preliminary assessing how well the CDR data from MTC can represent the distribution of the underlying population during the whole study period. The full 2011 Namibia Population and Housing Census data was obtained directly from the Namibia Statistics Agency. The official census date is 28 August but the actual enumeration exercise was undertaken over a period of about three weeks and ended on 15 September. The census-based population data depicts the spatial distribution of aggregated counts across 5475 enumeration areas (i.e. level 3 administrative unit), with a minimum, mean and maximum of 1, 386, and 5013 people per enumeration area, respectively. These data were subsequently further aggregated to administrative unit levels 2 and 1 to enable analysis of coarser spatial detail. Since aggregated counts at administrative unit levels only allow to assume that the population is evenly distributed throughout each spatial unit, we also used two gridded population datasets disaggregating the census-based population counts from enumeration areas into grid cells, without assuming equal distribution across space. WorldPop gridded population datasets (WorldPop, 2018), acquired for the years 2011, 2012 and 2013, are produced by using a top-down unconstrained Random Forest-based dasymetric approach (Stevens et al., 2015), which is relying on the use of a population density weighting layer, for redistributing census-based extrapolated estimates referring to each enumeration area into all ~100 m grid cells located within it—apart from the grid cells classified as water. The weighting layer is produced by using a population density response variable and a suite of ancillary data calculated at the administrative unit which are then used to fit a Random Forest model (Breiman, 2001) for predicting population density at the ~100 m grid cell level (Sorichetta et al., 2015).

Conversely, the GHS-POP gridded population dataset (Freire et al., 2016; Schiavina et al., 2019), acquired for 2014, are produced by using a top-down constrained approach which proportionally redistributes the same estimates used to produce the WorldPop datasets, from each enumeration area into its ~1 km grid cells according to their percentage of build-up area (Florczyk et al., 2019)—thus with the population of each enumeration area only redistributed with its grid cells that contain built settlements.

While the uncertainty and accuracy associated to each top-down gridded population distribution dataset may be difficult to quantify (Leyk et al., 2019), their use allow to consider two alternative modelled distributions, representing two extremes with respect to the population distribution derived from the counts aggregated at the enumeration are level, which can provide a more robust assessment of the population residing within the MTC network coverage.

Analysis and statistical tests

We explored and assessed four methods for distributing users into administrative unit levels: (1) Point to Polygon; (2) Voronois; (3) Tower Ranges; and (4) Adjusted Voronois. Each method's estimates of user density per unit was compared to census-derived population density figures, and Kendall's tau-b rank tests were used to determine the correlation between rankings. Seasonal and trend decomposition using Loess (STL) was subsequently used to analyse CDRs and identify patterns of seasonal user variation and investigate how the use of different distribution methods can impact upon identifiable seasonal patterns. Finally, administrative units sharing a similar seasonal behaviour were grouped into clusters.

Population coverage. Coverage of tower ranges across Namibia lies at approximately 30% of its area and is spatially incomplete, with coverage concentrating on urban areas and along the road network whilst not covering some remote areas. However, large portions of the land not covered by signals are very sparsely populated, so it is important to consider the tower coverage in terms of the population being covered and not just the area. Three population datasets were used to assess the proportion of the country's population which is covered by cellular towers in each year. The use of both gridded and non-gridded population datasets allows for a stronger assessment to be made about population coverage. Administrative unit level 3 2011 census data was intersected against the tower ranges for 2011 and had its counts proportionally represented, based on the area of each unit covered by tower ranges. Both WorldPop and GHS-POP data represent the number of people per grid cell and so the coverage of these data was calculated by summing the number of people in all grid cells contained within the tower ranges. WorldPop data was collected for and compared to each year the tower ranges represented. GHS-POP data was only available for 2014 and so was compared against the 2013 tower ranges.

Distribution methods

Each distribution method carries inherent benefits and drawbacks when being used to distribute users into administrative units. It is important to be aware of how well the redistributed user count numbers correlate with known population figures for the same area, in addition to the spatial coverage of each method. We perform each distribution method using models and python scripts in ArcGIS Pro 2.4.0.

The most basic method we implemented observes tower points which fall within a unit and then sums the counts from each point within it, shown in Fig. 3a. This is referred to as 'Point to Polygon' and has been used in a range of studies (Ihantamalala et al., 2018; Lai et al., 2019b; Schmid et al., 2017). This method is the least computationally demanding of those considered in this analysis. Due to the simplicity of the input data, Point to Polygon is unable to account for the signal range each cell tower has, meaning if a point falls within an administrative unit, all its counts will be assigned to it even if this cell tower provides service to people in neighbouring units; as is often the case when cell towers are located close to a boundary. Equally, this could limit the spatial extent of the method as every unit needs at least one cell tower to be located within it to receive a user count. This can vary between study sites but can lead to poor representation of parts of the study area characterised by low cell tower density, especially where units become small.

The second method being analysed requires additional information regarding the estimated signal range of each cell tower.

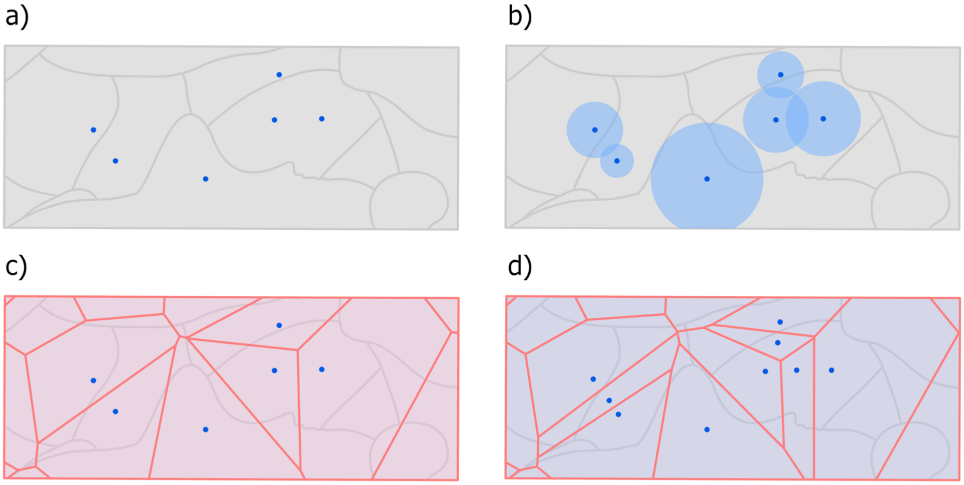


Fig. 3 Four methods of distributing CDR data into administrative units. a Point to Polygon, **b** Tower ranges, **c** Voronois, **d** Adjusted Voronois. Note: This figure was created using entirely fictitious data points, ranges and unit boundaries.

A buffer is created around each input point based on the known or estimated cell tower range (Fig. 3b). This represents the service area of each tower, which would theoretically translate to it being the most realistic method of distribution into administrative units. In our case, MTC provided approximate circular ranges for each tower. Tower Range counts are intersected against administrative units and have their user counts proportionally distributed across them, for example, if half of a tower range intersects unit X and the other half intersects unit Y, both units receive 50% of the counts. While this method has greater spatial coverage than Point to Polygon, there is no guarantee that it can provide total coverage of the study area; something that will be country and dataset specific.

The creation of Thiessen polygons or 'Voronoi' polygons around point data is thought to provide numerous benefits over simple Point to Polygon methodology. Using the cell tower point locations as inputs shown in Fig. 3c, Voronois form a continuous non-overlapping layer of polygons (Mercier and Baujard, 1997) representing the area of influence from each point to the nearest neighbouring point (Richter et al., 2019; Gautam et al., 2020). Where input points are more densely clustered, Voronoi polygons become smaller. This is typical of an urban area, contrasting rural settings which are likely to have fewer cell towers and much larger Voronoi polygons as a result. Voronoi polygons have been used in a wide range of studies to represent tower service areas (González et al., 2008; Gao et al., 2013; Xu et al., 2015; Zhao et al., 2016; Steele et al., 2017) and constitute the most frequently observed technique. User counts from the cell tower points are carried across to the corresponding Voronoi which allows them to be assigned to administrative boundaries proportionally, based upon the area for which they intersect. A key benefit of this method is its ability to provide 100% coverage across the study area, enabling all administrative units to be represented following count distribution. However, at the same time one needs to be aware that full coverage is an unrealistic representation which does not consider topology, radio frequency, or uninhabited areas like lakes. It is rather an artificial way of sub-setting the whole study area into continuous regions and has a tendency to overestimate (Koebe, 2020).

The final method is referred to as the 'Adjusted Voronoi' method. This method follows the methodology proposed by Ricciato et al. (2015) and attempts to enhance the standard Voronoi method through the introduction of new artificial input point data, based on tower range areas. Tower Ranges often exhibit overlap with the ranges of neighbouring towers, creating intersecting segments in the process. The Adjusted Voronoi method treats these segments as new polygons, then places an artificial cell tower point at their centre. The non-overlapping tower range polygon then has its cell tower point re-centred within its remaining area. Counts from the original tower ranges are proportionally distributed to the new set of points based on area. The new set of tower cell points is then used to create Voronoi polygons around them. The purpose of this is to create a higher number of input points to achieve a finer resolution Voronoi layer as depicted in Fig. 3d, which could make distribution into administrative units more accurate than the standard Voronois method.

Determining success of distribution methods. Using each distribution method, tower's daily average user counts per month were distributed into administrative units at three levels: Level 1, level 2 and level 3 for 2011, 2012 and 2013. We compare the 2011 CDR data to the 2011 Namibia Census Data (Namibia Statistics Agency, 2011) to determine how well the results of each method compare with reference data representing the same point in time. A comparison of mobile user counts versus census population counts was thought to be inappropriate in this case. Instead, it was decided to compare ranked administrative units and remove the influence of area to provide a directly comparable value across both sources. Census counts were converted to population density per administrative unit and ranked from highest to lowest. A three-month average of mean daily user counts from July to September 2011 was then used to create a user density value per unit for each distribution method at each level. A 3-month average was used to reduce the noise in the CDR data, whilst still matching the census collection date (from 28 August 2011 up to mid-September; Namibia Statistics Agency, 2011). As with the census units, a ranking number was assigned to each unit for each method.

We also explore if administrative units being classified as urban or rural can affect the observed correlations to the census in a separate test. All administrative units were considered for this, as above, with census and method derived values being split to include only urban or rural units at both level 2 and 3. Urban and rural classifications per administrative unit were informed by GHS-SMOD (Pesaresi et al., 2019; Florczyk et al., 2019) and GHS-POP gridded data (Schiavina et al., 2019; Freire et al., 2016), both acquired from the Joint Research Centre. As described in Florczyk et al. (2019), rural areas are an aggregation of L2 class topologies 10–13; suburban areas an aggregation of 21–23 and urban areas (city) are L2 class topology 30. Overall classification of an administrative units was then determined by population per class.

The distributions of both population and user density were heavily skewed, violating the normality criterion of the commonly used Pearson's correlation coefficient; instead Kendall's tau-b rank test was chosen to assess the correlation between user densities derived from the four distribution methods and census-derived population density. Kendall's tau-b is a non-parametric statistic which measures how well two sets of ranks for a pair of variables correspond to one another. A value of 1 indicates perfect agreement between two rankings. A value of 0 indicates that both ranks share no relationship, and −1 indicates the two rankings are totally opposite (Lasserre et al., 2011). This test was chosen over similar tests such as Spearman's Rank due to its ability to handle tied ranks (Javadi et al., 2016). All Kendall's Tau-b rank tests were performed using IBM SPSS Statistics 26. We consider the method exhibiting the highest correlation with the census to represent the most appropriate technique at the administrative level in question.

Decomposition analysis. STL was used to analyse administrative units' daily average user counts per month regarding seasonal behaviour and was carried out for the best and the worst performing CDR distribution method at administrative levels 2 and 3.

STL is a filtering procedure for decomposing a time series into trend, seasonal, and remainder components (Cleveland et al., 1990) as shown in Fig. 4. The algorithm implemented in R v3.6.1 finds the seasonal sub-series, for example the series of all January values, by taking the mean. It then removes these values and smooths the remainder to find the trend. In an iterative way, the overall level is subtracted from the seasonal component and added to the trend component. The residuals left from summing seasonal and trend form the remainder component (R Core Team, 2020).

Only the resulting seasonal component, which is no longer influenced by rising user counts due to STL's detrending algorithm, was used for further analysis. To enable the comparison of seasonal patterns between administrative units the seasonal component was

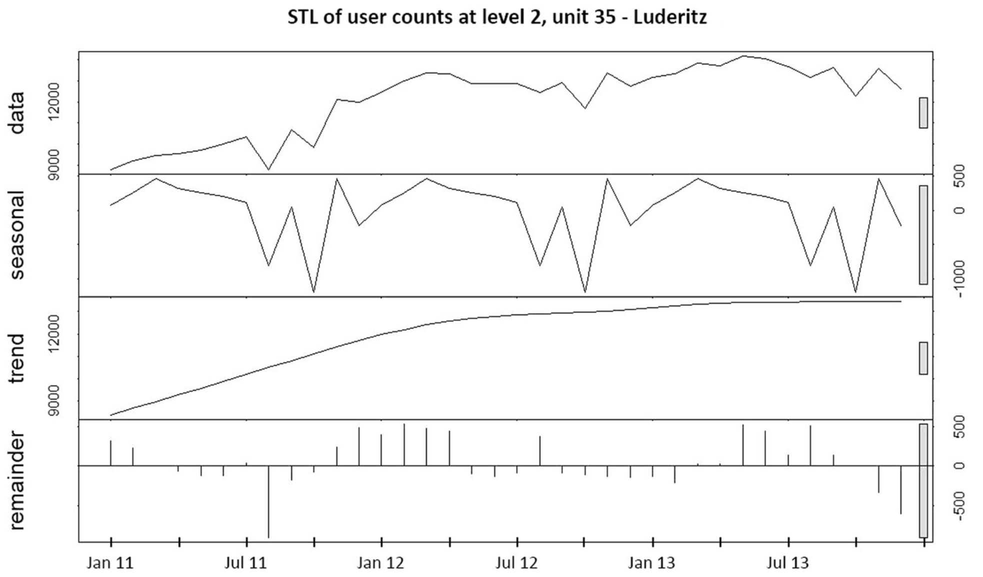


Fig. 4 Result of STL decomposition of the daily average of user counts per month for a level 2 administrative unit, Luderitz, using the corresponding best CDR distribution method, i.e. Tower Ranges. STL enables the separation of the seasonal component and the trend from CDR user counts (data) whilst leaving a remainder.

transformed into a seasonal proportion with an annual baseline. This calculation is in line with the initial processing of the CDR data into consistent annual tower sets but has the disadvantage of the resulting proportions potentially underestimating the population at the beginning of the year and overestimating at the end of the year due to general population and specific subscriber growth inherent in the annual user average used as denominator. The seasonal proportion informs what percentage of mobile phone users are in an administrative unit for a given month compared to the average number of people in that unit during the entire year. This seasonal proportion per administrative unit forms the basis for a clustering procedure.

Multivariate clustering. To analyse which administrative units share a similar seasonal behaviour, the seasonal proportions calculated in R were joined back to the spatial boundaries and used as input for a multivariate clustering tool in ArcGIS Pro 2.4.0. The tool was run with a *K*-medoids algorithm, which constructs nonspatial clusters and is more robust to noise and outliers than the more popular *K*-means algorithm (Environmental Systems Research Institute, 2020). Spatial clustering restricts clusters to be formed of neighbouring units only. In this application, there is little reason to believe that clusters must be spatially contiguous, hence a non-spatial clustering algorithm was chosen. A single cluster is formed from multiple administrative units based on the similarity of their seasonal user proportion alone; the location of the unit itself within Namibia or its proximity to other units is irrelevant.

Initially the tool was run without a restriction on the number of clusters as it was not known how many groups of similar seasonal behaviour to expect. The analysis of the computed pseudo-*F*-statistic for clustering solutions with 2–30 clusters gave an initial indication of how many clusters are sensible for the data. "The largest pseudo-*F*-statistic values indicate solutions that perform best at maximising both within-cluster similarities and between-cluster differences" (Environmental Systems Research Institute, 2020). Based on the three largest pseudo-*F*-statistic values we repeated the clustering with the associated number of clusters specified.

Table 2 Population coverage percentages in each year for three different population distribution datasets.

	Census (2011)	WorldPop (2011, 2012, 2013)	GHS-Pop (2014)
% coverage 2011	85.83	86.54	*86.32*
% coverage 2012	*86.41*	87.19	*86.91*
% coverage 2013	*87.03*	87.88	87.51

Italic values indicate a temporal mismatch between the population distribution dataset and the mobile phone tower ranges used for calculating the corresponding population coverage.

Table 3 Results of Kendall's tau-b rank test, comparing the order in which the administrative units are ranked by census-based population density figures against how the same units are ranked using user density figures derived from each CDR distribution method.

Administrative level	Point to Polygon	Voronois	Tower Range	Adjusted Voronois
Level 1	0.87	0.87	0.85	0.77
Level 2	0.81	0.86	0.89	0.79
Level 3	−0.08	0.55	0.60	0.52

Results

Population coverage of tower ranges. Mobile phone tower ranges cover most of Namibia's population with levels consistently higher than 85% regardless of the considered population distribution dataset (Table 2). The highest and lowest coverage are observed against the 2013 WorldPop dataset, at 87.8%, and the 2011 census-based data, at 85.8%, respectively.

Kendall's Tau-b rank. As seen in Table 3, high correlations are shown between the ranked results of each distribution method and the ranked census-based figures at Level 1, with Point to Polygon and Voronois tied for the highest score ($r = 0.87$). At level 2 more

Table 4 Results of the second Kendall's tau-b rank test, comparing the order in which administrative units are ranked by census-based population density figures against how the same units are ranked using user density figures derived from each CDR distribution method—only using administrative units covered by the corresponding distribution method.

Administrative Level	Point to Polygon	Voronois	Tower Range	Adjusted Voronois
Level 1	0.87	0.87	0.85	0.77
L1 units covered	13	13	13	13
Level 2	0.85	0.86	0.89	0.79
L2 units covered	105	107	107	107
Level 3	0.85	0.55	0.60	0.51
L3 units covered	520	5473	5099	5473
L3 % units lost	90.5	0	6.83	0

units are introduced and greater differences are observed. Results obtained using the Tower Ranges exhibit the highest correlation here (r = 0.89), with Adjusted Voronoi continuing to show the weakest results (Kendall's tau-b correlation, $r = 0.79$). At Level 3 correlations fall for all methods. Results obtained using Point to Polygon exhibit almost no relationship with the census-based figures at this level ($r = -0.08$), while results obtained using Tower Ranges continue to show the strongest correlation ($r = 0.60$). Both Voronois and Adjusted Voronois methods show strong results ($r = 0.55$ and $r = 0.52$, respectively).

At level three many units are not covered by some methods, leading to them all receiving a value of zero and subsequently being tied during the ranking process, thus impacting the correlation values. The second set of Kendall's tau-b correlation coefficients were calculated to judge the performance of each method without such tied ranks impacting the correlation values. This time, for each method, only administrative units which are covered by the specific method are included, and the census units are ranked again using only covered units. As shown in Table 4, this does not produce any change at level 1, as all units are covered. At level 2 there is no change for any method but Point to Polygon, which improves slightly ($r = 0.85$) while dropping 2 units. At level 3, Point to Polygon improves greatly ($r = 0.85$), although only 520 units are covered, leaving 90.5% of the total units at this level unaccounted for which renders this method inappropriate at this level. The correlation of Tower Ranges remains the same while dropping only 6.83% of units. Both Voronois and Adjusted Voronois see no change in correlation between Tables 3 and 4, as they provide 100% coverage across the study site.

A final set of Kendall's tau-b correlation coefficients were calculated to explore whether the classification of each administrative unit either urban or rural has any impact on how the ranked results of the different methods correlate to the ranked census-based figures. Due to small sample sizes, urban and suburban classes were combined for this analysis and level 1 was not included. It can be seen in Supplementary Table S2 that, when only rural units are considered, each distribution method is not as good at accurately ranking units at level 2, however, results obtained using Tower Ranges still exhibit the highest correlation at both level 2 and 3 ($r = 0.85$ and 0.46 respectively), while Point to Polygon and Adjusted Voronois continue to produce the worst results, with Point to Polygon showing, as in Table 3, a dramatic reduction in performance from level 1 to 3. The same pattern is observed when considering only the urban units, as results obtained using Tower Ranges once again yielded the highest correlation to the census-based figures at levels 2 and 3 ($r = 0.97$ and 0.35, respectively). Rankings of rural units at level 3 saw higher correlations to census-based values than city/suburban units at the same level in each method.

We assume that at each administrative level the methods producing results exhibiting the highest/lowest correlation with the census-based figures represent the most/least appropriate techniques, respectively. As demonstrated in Table 4, the number of units that each method can account for should also guide which methods to consider as appropriate. This means that, based on Tables 3 and 4, Tower Ranges are considered the best method at both levels 2 and 3 with high correlation and coverage. The least appropriate methods are shown to be Adjusted Voronois at level 2 and Point to Polygon at level 3. In the following stages of analysis, these best and worst methods will be compared against one another to assess their relative impacts on observable seasonal patterns. However, as Point to Polygon at level 3 covers <10% of administrative units and <10% of Namibia's population, a comparison between Tower Ranges and Point to Polygon would not provide relevant results at this level. Therefore, at level 3, Tower Ranges, representing the best method will be compared against the Adjusted Voronois representing the second-worst method.

Decomposition analysis. To compare the potential effect of using results from different distribution methods to inform subsequent analyses, an analysis of mobile phone users' seasonal mobility was carried out for both Tower Ranges and Adjusted Voronois at level 2 and 3, but not at level 1 as all methods performed very similarly at this level (Tables 3 and 4). Since both Tower Ranges and Adjusted Voronois make use of the tower ranges, Standard Voronois were included in the decomposition and clustering analysis in order to represent a method not requiring auxiliary data. Results are described in the Supplementary Information.

At level 2, STL was carried out to reveal the seasonal, trend and remainder components for each administrative unit: Fig. 4 shows the example of the level 2 administrative unit 'Luderitz' located in the South West of Namibia based on the results of the Tower Ranges method. The data clearly shows an increasing trend in daily averaged users counts per month starting at circa 9000 mobile phone users in January 2011 with an almost linear increase in users during the first year, before plateauing slowly from the beginning of 2012 onwards. The seasonal component is ranging between −1000 and +500 users across the 3 years and it can be clearly seen that the number of users decreases twice sharply in each year in August and in October.

To facilitate comparison between administrative units, seasonal components were subsequently converted to seasonal proportions. For both methods the largest seasonal increase was found in level 2 administrative unit 'Epukiro' in December 2011 (Tower Ranges = 57%, Adjusted Voronois = 56%). The largest seasonal decrease was found in level 2 administrative unit 'Karasburg' in August 2011 (Tower Ranges = −26%, Adjusted Voronois = −28%). See Fig. 5 for the location of these administrative units.

Figure 5 compares the largest seasonal increases and decreases between methods at level 2. For the majority of Namibia, the choice of distribution methods makes no difference when determining the month of the largest seasonal changes (green fill). For user increase and decrease, respectively, 79% and 78% of the administrative units match the month regardless of used method. Where there is no overlying hatching the percentage of change is also similar between methods (±3%). The few administrative units where the methods disagree (purple fill), vary depending on whether the increase or decrease is considered.

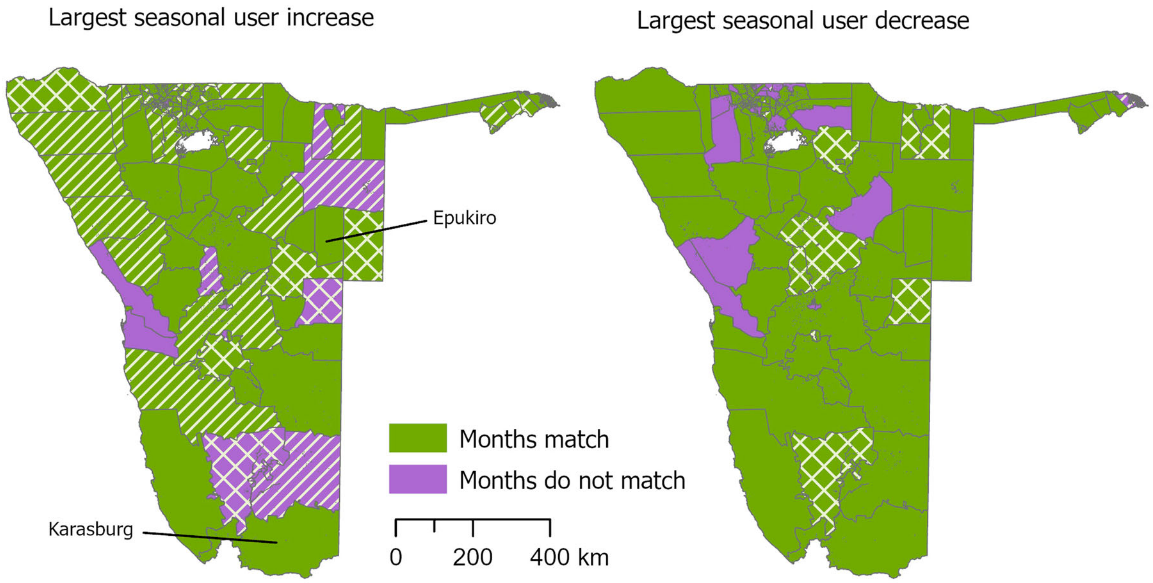


Fig. 5 Comparison between results from the Tower Range and Adjusted Voronois methods via the month with the largest user change at level 2. No hatch = both methods have similar values (±3%); simple hatch: Adjusted Voronois > Tower Ranges + 3%; cross hatch: Tower Ranges > Adjusted Voronois + 3%.

At level 3, the comparison is complicated by the fact that there are 374 administrative units which are not covered by a tower range and therefore were not considered for the decomposition of the Tower Range method. Another issue, affecting both Tower Ranges and Adjusted Voronois at level 3, is caused by an 'intersection effect': Distribution methods, like Tower Ranges, Voronois and Adjusted Voronois, require intersecting polygons with administrative units and are therefore prone to result, in some cases, in very small user numbers and in extreme cases in less than 1 user per administrative unit. This can happen when a Tower Range or Voronoi polygon intersects only a small portion of an administrative unit (sliver polygon) which will therefore receive only a small portion of the number of users originally assigned to the specific Tower Range or Voronoi polygon. The likelihood of this 'intersection effect' increases with the spatial scale (i.e. from level 1 to level 3) as intersections are more likely to result in sliver polygons due to more complex unit distribution. This can lead to abnormally large proportions where the sliver is produced by a tower built in a later year. This is because the user count (denominator) varies while the seasonal component (nominator) remains identical throughout the 3 years. While this 'intersection effect' had no impact on the results at level 2, its impact was obvious at level 3. In some cases, such 'intersection effect' resulted in seasonal proportions suggesting that up to 250k times more users would be present in a particular month compared to the annual average. To exclude these artefacts, and after considering the range of the seasonal proportions, we filtered out administrative units at level 3 with seasonal proportions >150% or < −150% to only keep units with data likely to represent actual change. This was also necessary to gain meaningful inputs for the following clustering procedure which would otherwise assign the affected units incorrectly to clusters. For the results from the Tower Range decomposition this meant that 36 administrative units were filtered out, whereas for Adjusted Voronois 55 units were disregarded. Units removed due to filtering or no overlapping tower ranges comprise 8% of the total units at level 3. This leaves a total of 5037 units for the comparison of the month of largest seasonal user change. Both Tower Ranges and Adjusted Voronois agree on the same month for the largest user change in 68% of units and disagree in 24% of units.

Seasonal decomposition was carried out for each administrative unit individually, therefore it cannot be expected that seasonal variations cancel each other out on a national level. However, we can assume that the majority of the CDR-based observed mobility is national travel rather than international, it is therefore reasonable to assume that mobility is approximately related to each other.

Multivariate clustering. Multivariate clustering at levels 2 and 3 for both methods, Tower Ranges and Adjusted Voronois, reveal that the ideal way to group administrative units for within-cluster similarity and between-cluster difference is to split units into two, three or four clusters with decreasing Pseudo-*F* values (see Supplementary Table S1).

When splitting administrative units into only two clusters, both methods lead to very similar outcomes. Figure 6 shows the results according to Tower Ranges with notable similarities in spatial distribution of clusters between administrative level 2 and 3. By relating the clusters back to the seasonal proportions and averaging them across each cluster, we can start to understand seasonal variation of user distribution within Namibia. With administrative units in cluster 1 (blue) gaining 20–30% of users in December whereas units in cluster 2 (red) are marked by a circa 10–15% user loss in August (see Fig. 6c).

Considering the example of four clusters at level 3 the difference in spatial distribution between the methods is very apparent (Fig. 7). Adjusted Voronois reveal a spatially homogeneous picture with 99% of the country being grouped either into cluster 1 (December peak) or 2 (August drop), an almost identical result to the grouping into 2 prescribed clusters (see Fig. 6b). The remaining 1% of units are allocated to cluster 3 and 4, which both show very strong seasonal patterns. Cluster 3 has a user decrease of circa 40% from January to August and a user increase of 50–80% between September and December. The same pattern was already observed for the corresponding level 2 administrative unit 'Karasburg', despite smaller changes at level 2 using Adjusted Voronois with 3 clusters (compare Supplementary Fig. S2) and using Tower Ranges with 4 clusters (Supplementary Fig. S3). Cluster 4 of Adjusted Voronois in Fig. 7 shows a new

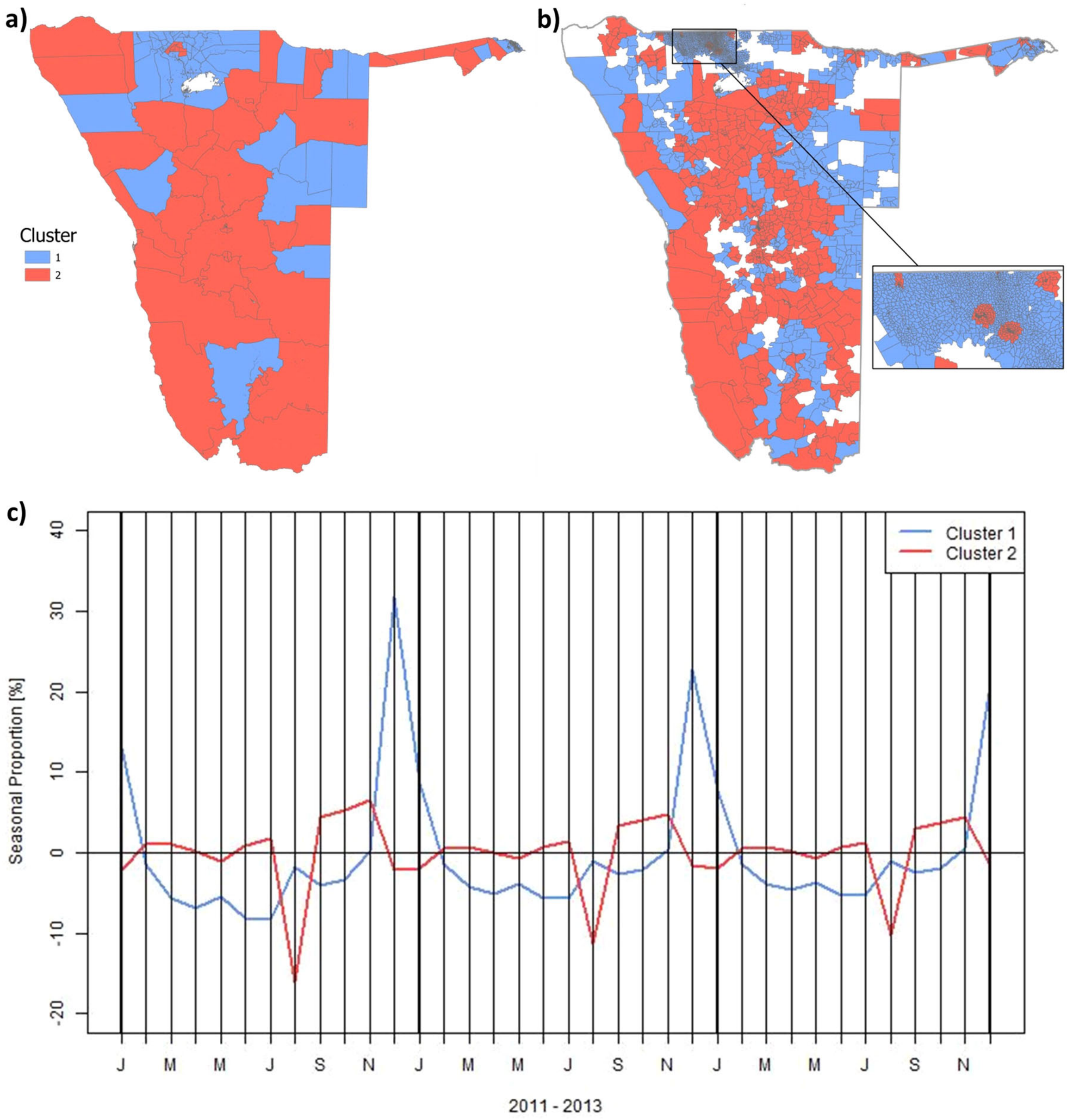


Fig. 6 Grouping of administrative units into two clusters using the Tower Range method. **a** At level 2, **b** at level 3, and **c** average seasonal user proportion per cluster at level 3. Administrative areas without fill colour have no data.

seasonal behaviour with user drops of 20–25% from January to April as well as in August, and user gains between 10% and 25% for the other months. Spatially, this cluster consists mainly of small urban administrative units in Windhoek but also several rural units in the North of Namibia. It is noticeable that while Adjusted Voronois' clustering results for three prescribed clusters (see Supplementary Fig. S4) shows one cluster marked by a drop in users in August and December, this behaviour is no longer visible when specifying four clusters.

Using Tower Ranges, the clusters appear spatially much more fragmented compared to Adjusted Voronois (compare Fig. 7a and b). Clusters 1 and 2 continue to show the already familiar patterns from previous clustering results, with the former cluster being characterised by the increase in users in December while the latter shows a drop in users in August. Cluster 3 shows a similar pattern to Adjusted Voronois but is more moderate with a user decrease from January until August ranging between 5% and 15% followed by a large increase up to ~30% in December. This is also similar to Cluster 4 using Tower Ranges at level 2 (Supplementary Fig. S3) which is likely due to the 3 Southern administrative units located within the equivalent level 2 unit 'Karasburg'. Cluster 4 in Fig. 7c shows partially similar patterns to cluster 1 but with an even more prominent user gain in December but no user change in August.

These results show that increasing the number of prescribed clusters leads to larger discrepancies between methods, both in terms of the spatial distribution of the clusters and the observed seasonal patterns. This behaviour was observed both at level 2 and 3, whereby the comparison of methods at level 3 is complicated by the fact that not all administrative units are covered by all methods (i.e. Tower Ranges do not cover the whole country) and the 'intersection effect' described in the Decomposition Analysis section. Comparisons of clustering results between methods for 3 clusters at level 2 can be found in Supplementary Fig. S2, for 4 clusters at level 2 in Supplementary Fig. S3 and for 3 clusters at level 3 in Supplementary Fig. S4. Clustering results based on the standard Voronoi method are also included in the Supplementary Information (Figs. S4–S7).

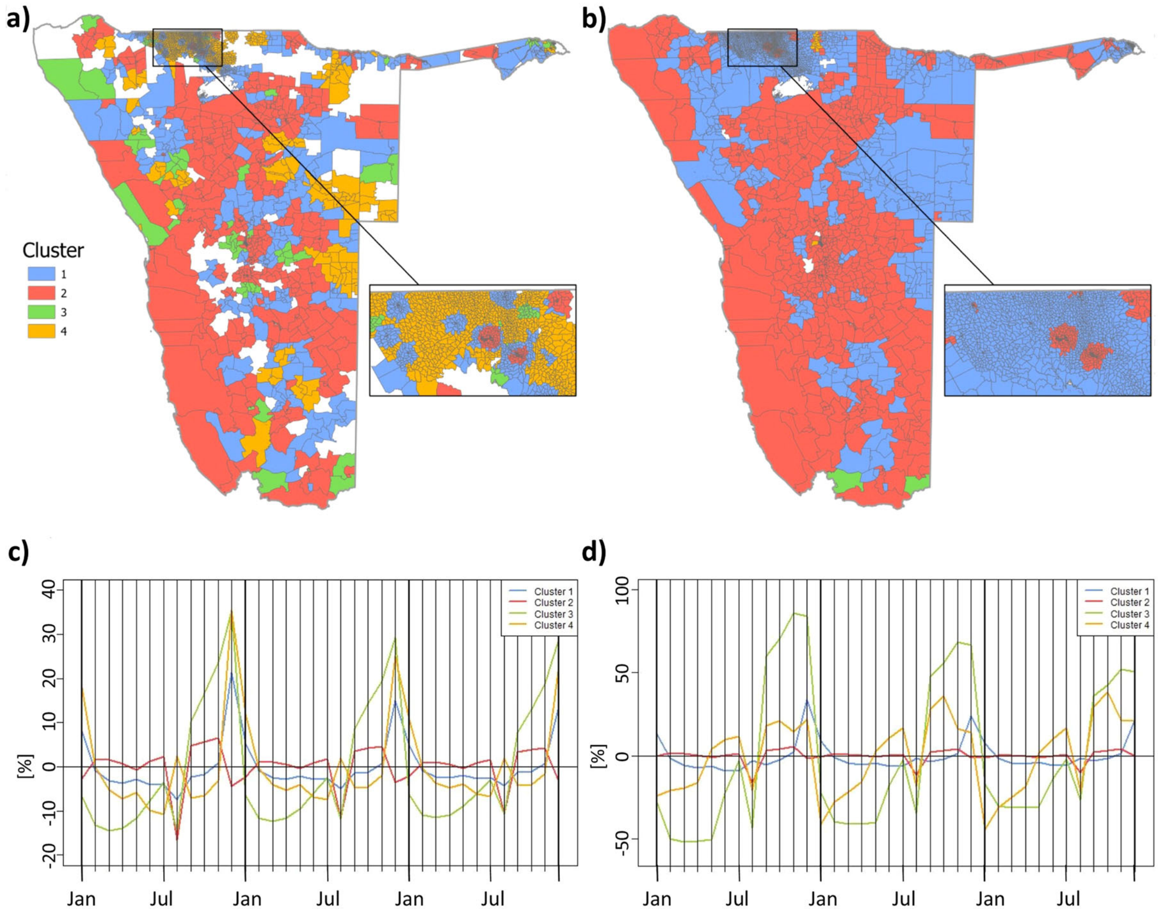


Fig. 7 Grouping of administrative units into four clusters at level 3. a Spatial distribution of clusters using Tower Ranges, **b** spatial distribution of clusters using Adjusted Voronois, **c** average seasonal user proportion per cluster from 2011 to 2013 using Tower Ranges, and **d** average seasonal user proportion per cluster from 2011 to 2013 using Adjusted Voronois. Administrative areas without fill colour have no data.

It is noticeable that seasonal proportions seem to attenuate over the years (see Figs. 6c and 7c, d). This is a side effect of the use of annual stable tower sets and the transformation of seasonal components to seasonal proportions as discussed above (see the section "Results—Decomposition analysis). Where MTC enlarged their network, an increasing number of towers is available in each year's tower set capturing an increasing number of users. The STL seasonal component is calculated in a way that it remains identical throughout the 3-year period. When converting seasonal components to seasonal proportions we divide the stable seasonal component (i.e. nominator) by the annually increasing user count (i.e. denominator), hence the artificial attenuation of the resulting proportions.

Despite the mentioned discrepancies, overall two major seasonal patterns can be identified across all distribution methods, administrative unit levels and most cluster configurations: (a) administrative units with a circa 15% decrease of mobile phone users in August and (b) administrative units with a 20–30% user increase in December. The only exception to this pattern is when Voronoi-based results are clustered into 3 groups at level 3 (see Supplementary Fig. S6).

Figure 8 shows the relationship between seasonal mobility and number of public and school holidays, with high mobility observed in August and December corresponding to the two months with the longest holidays.

Discussion

This study explores the use of different distribution methods at three different spatial scales and validates results based on census data. Results show high population coverage levels of MTC mobile phone tower ranges, above 85% for each source, mirroring the high mobile phone ownership of Namibian households (i.e. 89%; The Namibia Ministry of Health and Social Services and ICF International, 2014). As expected, the population coverage increases over time as a result of the expansion of MTC's network, which is noticeable in our data through the use of annual stable tower sets. The high agreement of coverage across different population distribution datasets strengthens the conclusions that can be drawn from the use of the CDR data, and suggests that it is appropriate to use such data for further analysis assessing the seasonal variation of population distribution in Namibia. Whilst the population coverage is shown to be high, it should still be recognised that a small portion of the population is missed. One might speculate that the missing ~12–15% of population live in rural parts of the country where cell coverage is either supplied by a different network provider or absent. The latter is supported by the 2013 demographic and health survey which found phone ownership of rural households being 14% lower than those of urban households (The Namibia Ministry of Health and Social Services and ICF International, 2014). This undersampling needs to be considered when results are used for decision and policy purposes.

The performance of distribution methods clearly varies with the spatial scale of the analysis; there is no single method consistently outperforming others, confirming that the spatial scale of the analysis needs to be considered to identify the best method. CDR data distributed through different methods translate into

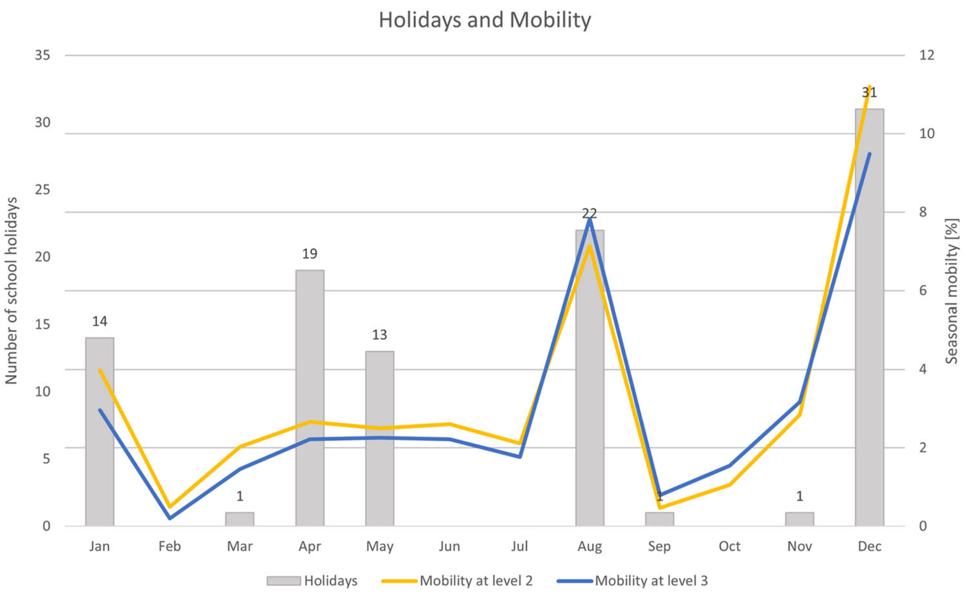


Fig. 8 Relationship between seasonal mobility and public and school holidays (obtained via Lai et al., 2020b; see "Data availability" section). Mobility was averaged per month from 2011 to 2013 across all administrative units.

input data of varying accuracy which therefore impacts subsequent analysis as shown by the clustering results.

At level 1, there is little difference observed between each method's correlation with census derived population density rankings (Tables 3 and 4). This is likely a result of units at this level being large in area and few in number, with tower density therefore being relatively high per administrative unit. Point to Polygon presents itself as being an effective method at this scale, in addition to being easy to implement and requiring no additional information. The three other methodologies also perform well, however the additional steps and information required to implement them are shown to provide no discernible benefits at this spatial scale.

With 107 units, the tower density per administrative unit at level 2 is much lower. This results in more variation being observed in correlations with Point to Polygon and Voronoi methods both exhibiting weaker correlations, while Tower Ranges and Adjusted Voronois show slight improvements. Tower Range methodology performs best; the ability to carry out proportional distribution, coupled with more accurate information regarding each tower's service area enables this method to better handle increased boundary complexity and number of units. While the correlation fell slightly at level 2, Voronois appear to benefit from their ability to proportionally distribute users based on the area of intersection, in addition to maintaining distribution to all units. These factors could be what leads this method to outperform Point to Polygon at this level, having not done so at level 1. Using Voronoi polygons as an approximate cellular range for each point is however shown to be far less effective than using accurate Tower Range information. The drawbacks of Point to Polygon begin to emerge, with an increased number of units leading to greater boundary complexity. The simple assignment of points to units without accounting for proximity to neighbouring units results in misplacing users who would be served by the tower in practice. Moreover, there are not enough input points to cover all available units and so this method is limited by the assumption that units without an input point are entirely devoid of users. Adjusted Voronois display a slight improvement between levels 1 and 2, although it still shows the lowest correlation of all methods. Despite increased variation in correlation, clustering results at level 2 are largely similar: Tower Ranges, Voronois and Adjusted Voronois all agree on an ideal split into two clusters with very similar spatial distribution and seasonal patterns. When forcing a split into more clusters, discrepancies in spatial distribution and seasonal patterns emerge.

Level 3 represents the biggest challenge for these methods; with over 5000 units to account for, tower density is very low and more than 90% of units are no longer represented by a tower at all without the help of estimated ranges. Tower Ranges continue to outperform the other methods, however, cannot account for 6.83% of units which do not fall within them. Both Voronoi methods perform well considering the number and density of units, however, the inaccuracies of using Voronois as a proxy for service area seemingly become more apparent with reduced correlation values. Point to Polygon shows no correlation to census rankings as only a small fraction of units is covered. At level 3, clustering results have more variation than at level 2, but are still similar when clustering into only two groups. As already observed at level 2, higher cluster numbers lead to an increase in variation.

The observed correlations to census ranking also hold true when stratifying administrative units into urban and rural areas. Unexpectedly, rural units at level 3 across all methods had a higher correlation with the census than city/suburban administrative unit at the same level, this is possibly because city/suburban units are typically much smaller in size than rural ones, making it harder to distribute user counts from each method accurately, as shown in Tables 1 and 2.

Koebe (2020) suggests that more precise distribution methods do not automatically lead to more precise outcomes. Our results show that this is true only at low administrative-unit level with high tower density, in our case at level 1, but not at level 2 and 3 where Tower Ranges with their more accurate representation of the service area outperform the other methods. Koebe's work in Senegal is done at high administrative-unit level (i.e. level 4 which splits the country into 431 communes), which initially seems to contradict our findings. However, number and distribution of cell towers of the network provider in Senegal are favourable with nearly all units covered by at least one of the 1666 towers leading to a tower density of ~8 towers per 1000 km^2. This is substantially higher than in Namibia where tower density is <1 tower per 1000 km^2 for most level 1 units and only rises to a maximum of 4.6 (see Fig. 2). In this context, Koebe's work which aggregates data to level 4 is therefore more comparable to our level 1 data and results.

With regards to Namibia, the Adjusted Voronoi method proposed by Ricciato et al. (2015) performed worse than the standard Voronoi method across all spatial scales. Whether this is due to the irregular distribution of cell towers across the country or a more general result of this method, which was developed on artificial data, remains open. Our results show that if estimates of tower service areas are available, better performance can be obtained when they are used directly as Tower Ranges and not to adjust Voronois.

Our results also show that CDR data, aggregated to users per tower and month, are suitable to derive seasonal mobility through the use of STL; exact information of individual's locations at various time stamps is therefore not always necessary to carry out mobility analysis. STL in combination with multivariate clustering delivers valuable insight into which areas of a country share a similar seasonal variation in mobile phone users. We found that the higher the number of prescribed clusters, the more discrepancies between methods. This is to be expected as the sheer number of clustering options increases along with the number of clusters used. Despite the discrepancies, two major seasonal patterns can be identified across all distribution methods, administrative levels and most cluster numbers: (a) areas with a circa 15% decrease of mobile phone users in August and (b) areas with a 20–30% user increase in December. Pseudo-*F* values clearly favour clustering into these two groups. Cluster (b) seconds the findings of zu Erbach-Schoenberg et al. (2016) who analysed seasonal mobility using the relatively small scale of health districts. With 33 health districts in Namibia, they are closest to administrative unit level 1 in terms of spatial scale. To gain further insights into seasonal movements we can prescribe a higher cluster number whilst using the best performing method at the highest spatial resolution; Tower Ranges at level 3 (Fig. 7a and c). However, these results have to be considered with caution due to their lower Pseudo-*F* values.

In general, the highest average mobility is observed in August and December (Fig. 8), the two months with the longest holidays. With primary school net enrolment of 91% between 2010 and 2014 (UNICEF, 2016) and over 90% of Namibia's population identifying themselves as Christians (Horn, 2008), it is likely that some of the seasonal mobility patterns detected are due to people travelling during school holidays to celebrate Christian festivities with relatives and friends. Similarly, mobility is increasing to a lesser extent during the shorter holidays in April and May, which can partially be linked to religious events like Easter and Ascension Day but also to other non-religious public holidays (e.g. half-term, Africa Day, Day of the African Child). However, other changes in phone usage that are entirely unrelated to mobility could also have an impact, for example children being allowed to stay up later during holidays and/or using their phones more, or adults being less active over the phone while spending time with their family.

The level 2 administrative unit 'Karasburg' located in the South of Namibia (see Fig. 5) stands out in this analysis as it is repeatedly listed as a separate cluster and is characterised by the largest user decrease between January and August (>20%) and the largest user increase between September and December (>30%). Under closer examination we found that this is likely to be caused by the presence of grape farms using water from the Orange River as source for irrigation. The special combination of water availability and hot climate in this part of Namibia allows table grapes here to be harvested at the end of the year, when they are most vulnerable to frost elsewhere in the world making them a lucrative export product to the European market (Hoffman, 2008). The growing of grapes is very labour intensive, especially during harvest, which requires additional workers. Based on a combination of internet searches and visual analysis of satellite images there seem to be 3 grape farms located along the Orange River: Komsberg Farm, Sonop Farm, and the Orange River Irrigation Project at Aussenkehr, with the latter being by far the largest. In 2011, Aussenkehr had close to 30,000 residents with an additional 7000 harvest workers for 3–4 months a year (The Namibian, 2011). This large in- and outflow of seasonal workers is also what the CDR data show clearly with a large increase of users from September to December during the grape harvest and a large user decrease between January and August when there is less work to do in the vineyards. In the underdeveloped region of Karasburg, where all three farms are located, the grape industry is the largest employer; to produce and export 1000 tonnes of table grapes it is estimated that 300 new permanent and 600 seasonal jobs are created (Hoffman, 2008).

Looking at cluster 3 in Fig. 7b, we can see the level 3 units in which Aussenkehr and Komsberg Farm are located. At this larger scale users of this cluster almost double between September and December (80%). The administrative unit of Sonop Farm is not included in this cluster. We can speculate that this is because it is a smaller farm and because it is growing a wider range of fruits and vegetables like tomatoes and peppers with different growing and harvest cycles to grapes (Sonop Farms, 2020), therefore having more permanent and less seasonal workers.

With agriculture and herding forming large parts of Namibia's economy (The Namibia Ministry of Health and Social Services and ICF International, 2014), a link between agricultural activity and seasonal mobility is no surprise. Further investigation is necessary to see if and to what extent the seasonal behaviour of other clusters can be connected to agriculture and whether they correlate better than holidays. Zufiria et al. (2018), who used more detailed CDR data were able to quantify relationships not only between mobility and the agricultural calendar, but general economic activity as well as rainfall. It would be interesting to see whether similar relationships can be established in Namibia where the agricultural sector supports directly or indirectly over 70% of the population (Namibia High Commission London, 2022).

The results of this study are affected by a number of methodological limitations which need to be considered when using them for informing and supporting decision and policy makers. Firstly, using the modelled distribution of mobile phone users as a proxy for the distribution of the underlying population is a simplification and it needs to be considered that the observed changes in mobile phone usage may not be a function of mobility of users alone. Other changes in phone usage, for example a different usage behaviour during holidays, could also have an impact on user count fluctuations. Secondly, people may have multiple mobile devices/SIMs which would artificially inflate the observed CDR-based mobility. Thirdly, CDR data are biased towards particular demographics, not only ownership of mobile phones but also the likelihood and frequency of their usage will vary between demographic groups. While growth in cell phone adoption rate is corrected for by the detrending algorithm of STL, when using annual baselines for the calculation of seasonal proportions we effectively introduce growth rates back into the data. In combination with distribution methods using intersection, this can lead to unrealistically large seasonal proportions. The likelihood of this happening increases with larger spatial scales (i.e. level 3) and potentially requires post processing, for example in the form of filtering based on a data identified threshold, in our case ±150%. In this study, this shortcoming was accepted so as not to lose valid input data and therefore obtain the greatest coverage possible, which is of particular importance in a country like Namibia where tower density is relatively low, especially in rural areas. The decision on which approach to use must also consider the desired outcomes. For example, in settings with higher tower densities or if the scope is to estimate precise population numbers (like in the works of Zufiria et al., 2018 or Salat et al., 2020), this problem can be easily avoided by using a consistent set of towers for the entire study period. Finally, assuming rural towers are also more likely to be inactive for longer periods due to more difficult access for repairs, the likelihood of undersampling in the countryside increases. This must be considered when formulating policies especially as the majority of Namibia's population is still living in rural settings during the timeframe of this study.

Conclusions

This study, like many others, shows that CDR data can be used for estimating seasonal mobility and the associated changes in population distribution at the subnational level (Deville et al., 2014; Ricciato et al., 2015; zu Erbach-Schoenberg et al., 2016; Lai et al., 2019b). However, it also shows that in a country like Namibia, where mobile phone tower coverage is limited in rural areas, it is important and recommended to investigate both what proportion of population is covered by the tower ranges and, most importantly, which is the best spatial method to distribute the mobile phone users from towers to administrative units in order to best represent the underlying population distribution and inform any further work or seasonal analysis.

The method used to distribute CDRs into areas of interest, e.g. administrative units or health facility catchments, can impact upon the accuracy to which counts are allocated to them and therefore needs to be considered carefully at the outset. This study shows that the performance of each method is greatly influenced by the spatial scale and resulting tower density at which they are analysed. We conclude that Tower Ranges represent the best distribution method when working at high administrative-unit level with relatively low tower density per unit, i.e. levels 2 and 3; benefiting greatly from the more accurate coverage estimation which is absent from other methods. It is also shown that Voronois can provide a useful alternative if tower range information is unavailable. In scenarios where there are enough towers to cover all units, which is typically the case at low administrative-unit level with relative high tower density per unit (i.e. level 1, see Fig. 2), Point to Polygon performs well and serves as a viable, easy to implement option.

With regards to Namibia, the Adjusted Voronoi method proposed by Ricciato et al. (2015) performed worse than the standard Voronoi method across all spatial scales. Whether this is due to the irregular distribution of cell towers across the country or a more general result of this method, which was developed on artificial data, remains a question for future research using actual CDRs from different countries and settings.

STL has shown to be a valuable way of revealing seasonal variation in mobile phone user distribution. The example of 'Karasburg', an important region for the growing of table grapes, has proven that CDR data has the ability to show the extent of seasonal mobility of harvest workers which needs to be considered when planning new infrastructures. Similarly, the grouping of administrative units by their seasonal behaviour facilitates the discovery of patterns, for example in identifying areas with an increased population around the Christmas period, which consideration should be given to when discussing policies for resource allocation. Similar results, showing the importance of accounting for seasonal movements when calculating disease incidence were shown in zu Erbach-Schoenberg et al. (2016). Awareness of seasonal mobility plays also an important role in the field of epidemiology and can aid identification of seasonally varying transmission hotspots of infectious diseases as well as inform the targeting of suitable interventions, for example, the spatio-temporal distribution of bed nets to minimise malaria transmission (Tatem et al., 2014; Ruktanonchai et al., 2016). In addition, the underlying CDR distribution method can have a major impact on subsequent analysis as shown in this study on the example of multivariate clustering (Fig. 7). Seasonal change as a proportion of the usual annual population is an easy to communicate metric which can be beneficial for supporting specific policies/actions.

Future research should concentrate on validating the results regarding the benefit of Tower Ranges over Voronoi based and Point to Polygon methods in different countries and settings. This would be especially interesting in countries with near-complete tower penetration where Point to Polygon might reveal consistently good results across all spatial scales without the issue of missing out spatial units. To make the results more widely usable, it would be a useful next step to try and quantify a threshold of tower density at which different distribution methods can be recommended. Additional factors which are likely to play a role, for example, the percentage of units covered by cell towers and boundary complexity, should also be further examined.

Data availability

The Call Data Record dataset analysed during the current study is not publicly available since that would compromise the agreement with the mobile phone operator that made the data available for research, but information about the process of requesting access to the mobile phone data that support the findings of this study are available from the corresponding author on reasonable request.

Population data used in this study are available via the following sources:

- GHS-POP Data

Schiavina M, Freire S, MacManus K (2019) GHS population grid multitemporal (1975, 1990, 2000, 2015) R2019A. European Commission, Joint Research Centre (JRC) https://doi.org/10.2905/42E8BE89-54FF-464E-BE7B-BF9E64DA5218 PID: http://data.europa.eu/89h/0c6b9751-a71f-4062-830b-43c9f432370f Accessed: 30 Apr 2020.

Concept & Methodology:

Freire, Sergio; MacManus, Kytt; Pesaresi, Martino; Doxsey-Whitfield, Erin; Mills, Jane (2016): Development of new open and free multi-temporal global population grids at 250 m resolution. Geospatial Data in a Changing World; Association of Geographic Information Laboratories in Europe (AGILE). AGILE 2016.

- GHS-SMOD Data

Pesaresi, Martino; Florczyk, Aneta; Schiavina, Marcello; Melchiorri, Michele; Maffenini, Luca (2019): GHS settlement grid, updated and refined REGIO model 2014 in application to GHS-BUILT R2018A and GHS-POP R2019A, multitemporal (1975-1990-2000-2015), R2019A. European Commission, Joint Research Centre (JRC) [Dataset] https://doi.org/10.2905/42E8BE89-54FF-464E-BE7B-BF9E64DA5218 PID: http://data.europa.eu/89h/42e8be89-54ff-464e-be7b-bf9e64da5218.

Concept & Methodology:

Florczyk, Aneta J.; Corbane, Christina; Ehrlich, Daniele; Freire, Sergio; Kemper, Thomas; Maffenini, Luca; Melchiorri, Michele; Pesaresi, Martino; Politis, Panagiotis; Schiavina, Marcello; Sabo, Filip; Zanchetta, Luigi (2019): GHSL Data Package 2019, EUR 29788 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-13187-8 https://doi.org/10.2760/0726, JRC 117104.

- Census Data

Namibia Statistics Agency, Government of the Republic of Namibia. Namibia Population and Housing Census 2011. Received: June 2015.

- WorldPop Population Estimates, 100 m

WorldPop (www.worldpop.org—School of Geography and Environmental Science, University of Southampton; Department of Geography and Geosciences, University of Louisville; Departement de Geographie, Universite de Namur) and Center for International Earth Science Information Network (CIESIN), Columbia University (2018). Global High Resolution Population Denominators Project—Funded by The Bill and Melinda Gates Foundation (OPP1134076). https://doi.org/10.5258/SOTON/WP00660.

- WorldPop Population Density Estimates, 1 km

WorldPop (www.worldpop.org—School of Geography and Environmental Science, University of Southampton; Department of Geography and Geosciences, University of Louisville; Departement de Geographie, Universite de Namur) and Center for International Earth Science Information Network (CIESIN), Columbia University (2018). Global High Resolution Population Denominators Project - Funded by The Bill and Melinda Gates Foundation (OPP1134076). https://doi.org/10.5258/SOTON/WP00675.

The administrative unit boundaries are available from:

- WorldPop Namibia Shapefiles

WorldPop (www.worldpop.org—School of Geography and Environmental Science, University of Southampton; Department of Geography and Geosciences, University of Louisville; Departement de Geographie, Universite de Namur) and Center for International Earth Science Information Network (CIESIN), Columbia University (2018). Global High Resolution Population Denominators Project—Funded by The Bill and Melinda Gates Foundation (OPP1134076). https://doi.org/10.5258/SOTON/WP00651.

- GADM Namibia Shapefiles

Global Administrative Areas (2022). GADM database of Global Administrative Areas, version 4.0. URL: www.gadm.org.

The global and school holidays dataset is available from:

- Lai S, Sorichetta A and WorldPop (2020) Global Public and School Holidays 2010-2019. Mapping seasonal denominator dynamics in low- and middle-income settings. https://doi.org/10.5258/SOTON/WP00691.

Received: 25 February 2021; Accepted: 29 June 2022;
Published online: 28 July 2022

References

Breiman L (2001) Random forests. Mach Learn 45:5–32

Cleveland RB, Cleveland WS, McRae JE, Terpenning I (1990) STL: a seasonal-trend decomposition procedure based on loess. J Off Stat 6(1):3–73

Davis KF, D'Odorico P, Laio F, Ridolfi L (2013) Global spatio-temporal patterns in human migration: a complex network perspective. PLoS ONE 8(1):e53723

Deville P, Linard C, Martin S, Gilbert M, Stevens FR, Gaughan AE, Blondel VD, Tatem AJ (2014) Dynamic population mapping using mobile phone data. Proc Natl Acad Sci USA 111(45):15888–15893. https://doi.org/10.1073/pnas.1408439111

Environmental Systems Research Institute (ESRI) (2020) ArcGIS pro tool help multivariate clustering (spatial statistics). https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/multivariate-clustering.htm. Accessed 16 Nov 2020

Fernández IC, Wu J (2018) A GIS-based framework to identify priority areas for urban environmental inequity mitigation and its application in Santiago de Chile. Appl Geogr 94:213–222. https://doi.org/10.1016/j.apgeog.2018.03.019

Florczyk AJ, Corbane C, Ehrlich D, Freire S, Kemper T, Maffenini L, Melchiorri M, Pesaresi M, Politis P, Schiavina M, Sabo F (2019) GHSL data package 2019. European Commission Joint Research Center, 29788

Freire S, MacManus K, Pesaresi M, Doxsey-Whitfield E, Mills J (2016) Development of new open and free multi-temporal global population grids at 250 m resolution. Geospatial Data in a Changing World; Association of Geographic Information Laboratories in Europe (AGILE). AGILE 2016

Frias-Martinez E, Williamson G, Frias-Martinez V (2011) An agent-based model of epidemic spread using human mobility and social network information. In: Conference Proceediengs of 2011 IEEE third international conference on privacy, security, risk and trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), IEEE, Boston, USA, 9–11 October 2011, pp. 57–64, https://doi.org/10.1109/PASSAT/SocialCom.2011.142

Gao S, Liu Y, Wang Y, Ma X (2013) Discovering spatial interaction communities from mobile phone data. Trans GIS 17(3):463–481

Gaughan AE, Stevens FR, Huang Z, Nieves JJ, Sorichetta A, Lai S, Ye X, Linard C, Hornby GM, Hay SI, Yu H (2016) Spatiotemporal patterns of population in mainland China, 1990 to 2010. Sci Data 3:160005. https://doi.org/10.1038/sdata.2016.5

Gautam S, Brema J, Dhasarathan R (2020) Spatio-temporal estimates of solid waste disposal in an urban city of India: a remote sensing and GIS approach. Environ Technol Innov 18:100650

González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. Nature 453(7196):779–82. https://doi.org/10.1038/nature06958

Gwitira I, Murwira A, Masocha M, Zengeya FM, Shekede MD, Chirenda J, Tinago W, Mberikunashe J, Masendu R (2019) GIS-based stratification of malaria risk zones for Zimbabwe. Geocarto Int 34(11):1163–1176

Hoffman J (2008) Greening the Namibian Desert: an African Success Story. The South African Institute of International Affairs. African perspectives. Global insights. https://saiia.org.za/research/greening-the-namibian-desert-an-african-success-story/. Accessed 16 Nov 2020

HornN (2008) Religion and human rights in Namibia. Afr Hum Rights Law J 8:409–431. http://www.ahrlj.up.ac.za/images/ahrlj/2008/ahrlj_vol8_no2_2008_nico_horn.pdf. Accessed 16 Nov 2020

Ihantamalala FA, Herbreteau V, Rakotoarimanana FM, Rakotondramanga JM, Cauchemez S, Rahoilijaona B, Pennober G, Buckee CO, Rogier C, Metcalf CJE, Wesolowski A (2018) Estimating sources and sinks of malaria parasites in Madagascar. Nat Commun 9(1):1–8

Javadi MS, Dahl M, Pettersson M (2016) Change detection in aerial images using a Kendall's TAU distance pattern correlation. In: 2016 6th European Workshop on Visual Information Processing (EUVIP). IEEE, Marseille, France, 25–27 October 2016, pp. 1–6, https://doi.org/10.1109/EUVIP.2016.7764604

Jiang S, Ferreira J, Gonzalez MC (2017) Activity-based human mobility patterns inferred from mobile phone data: a case study of Singapore. IEEE Trans Big Data 3(2):208–219

Kanasugi H, Sekimoto Y, Kurokawa M, Watanabe T, Muramatsu S, Shibasaki R (2013) Spatiotemporal route estimation consistent with human mobility using cellular network data. In: Conference Proceedings of 2013 IEEE international conference on Pervasive Computing and Communications Workshops (PERCOM Workshops). IEEE, San Diego, USA, 18–22 March 2013, pp. 267–272, https://doi.org/10.1109/PerComW.2013.6529493

Koebe T (2020) Better coverage, better outcomes? Mapping mobile network data to official statistics using satellite imagery and radio propagation modelling. PLoS ONE 15(11):e0241981. https://doi.org/10.1371/journal.pone.0241981

Kraemer MU, Reiner RC, Brady OJ, Messina JP, Gilbert M, Pigott DM, Yi D, Johnson K, Earl L, Marczak LB, Shirude S (2019) Past and future spread of the arbovirus vectors *Aedes aegypti* and *Aedes albopictus*. Nat Microbiol 4(5):854–63. https://doi.org/10.1038/s41564-019-0376-y

Lai S, Bogoch I, Ruktanonchai N, Watts A, Lu X, Yang W, Yu H, Khan K, Tatem AJ (2020a) Assessing spread risk of Wuhan novel coronavirus within and beyond China, January–April 2020: a travel network-based modelling study. Preprint at medRxiv https://doi.org/10.1101/2020.02.04.20020479

Lai S, Farnham A, Ruktanonchai, NW, Tatem, AJ (2019a) Measuring mobility, disease connectivity and individual risk: a review of using mobile phone data and mHealth for travel medicine. J Travel Med 26(3), https://doi.org/10.1093/jtm/taz019

Lai S, Ruktanonchai NW, Zhou L, Prosper O, Luo W, Floyd JR, Wesolowski A, Santillana M, Zhang C, Du X, Yu H (2020b) Effect of non-pharmaceutical interventions to contain COVID-19 in China. Nature 585:410–413. https://doi.org/10.1038/s41586-020-2293-x

Lai S, zu Erbach-Schoenberg E, Pezzulo C, Ruktanonchai NW, Sorichetta A, Steele J, Li T, Dooley CA, Tatem AJ (2019b) Exploring the use of mobile phone data for national migration statistics. Palgrave Commun 5(1):1–10. https://doi.org/10.1057/s41599-019-0242-9

Lasserre JF, Pop-Ciutrila IS, Colosi HA (2011) A comparison between a new visual method of colour matching by intraoral camera and conventional visual and spectrometric methods. J Dent 39:e29–e36

Lenormand M, Louail T, Cantú-Ros OG, Picornell M, Herranz R, Arias JM, Barthelemy M, San Miguel M, Ramasco JJ (2015) Influence of sociodemographic characteristics on human mobility. Sci Rep 5:10075

Linard C, Gilbert M, Snow RW, Noor AM, Tatem AJ (2012) Population distribution, settlement patterns and accessibility across Africa in 2010. PLoS ONE 7(2):e31743. https://doi.org/10.1371/journal.pone.0031743

Leyk S, Gaughan AE, Adamo SB, de Sherbinin A, Balk D, Freire S, Pesaresi M (2019) The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. Earth Syst Sci Data 11(3):1385–1409

Mercier F, Baujard O (1997) Voronoi diagrams to model forest dynamics in French Guiana. Paper presented at the second annual conference of GeoComputation '97 & SIRC '97. University of Otago, New Zealand, 26–29 August 1997

Midzi N, Kavhu B, Manangazira P, Phiri I, Mutambu SL, Tshuma C, Chimbari MJ, Munyati S, Midzi SM, Charimari L, Ncube A (2018) Inclusion of edaphic predictors for enhancement of models to determine distribution of soil-transmitted helminths: the case of Zimbabwe. Parasites Vectors 11(1):1–13

Namibia High Commission London (2022) About Namibia's economy—agriculture. https://www.namibiahc.org.uk/economy.php. Accessed 25 Apr 2022

Namibia Statistics Agency (2011) Government of the Republic of Namibia. Namibia population and housing census 2011. Namibia Statistics Agency (Received: June 2015)

Namibia Statistics Agency (2013) Namibia 2011 Census Atlas. Windhoek, Namibia. https://nsa.org.na/page/publications. Accessed 25 Apr 2020

Namibia Statistics Agency (2015) Namibia 2011 Census Migration Report. Windhoek, Namibia. https://nsa.nsa.org.na/wp-content/uploads/2021/09/p19dmqq344hnc6ji1cioctа1eg21.pdf. Accessed 31 Mar 2022

Palmer R, Espenshade TJ, Bartumeus F, Chung CY, Ozgencil NE, Li K (2013) New approaches to human mobility: Using mobile phones for demographic research. Demography 50(3):1105–1128

Pendleton W, Crush J, Nickanor N (2014) Migrant Windhoek: Rural–urban migration and food security in Namibia. Urban Forum 25(2):191–205 https://doi.org/10.1007/s12132-014-9220-x

Poel M, Meyer ET, Schroeder R (2018) Big data for policymaking: great expectations, but with limited progress? Policy Internet 10(3):347–367

R Core Team (2020) A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.rdocumentation.org/packages/stats/versions/3.6.1/topics/stl. Accessed 16 Nov 2020

Ricciato F, Widhalm P, Craglia M, Pantisano F (2015) Estimating population density distribution from network-based mobile phone data. Publications Office of the European Union, European Commission, Joint Research Centre (JRC) Technical Report, Luxembourg, 10.2788/162414

Ricciato F, Widhalm P, Pantisano F, Craglia M (2017) Beyond the "single-operator, CDR-only" paradigm: an interoperable framework for mobile phone network data analyses and population density estimation. Pervasive Mob Comput 35:65–82. https://doi.org/10.1016/j.pmcj.2016.04.009

Ricciato F, Coluccia A (Preprint) On the estimation of spatial density from mobile network operator data. IEEE Transactions on Mobile Computing. https://doi.org/10.1109/TMC.2021.3134561

Richter A, Ng KTW, Karimi N (2019) A data driven technique applying GIS, and remote sensing to rank locations for waste disposal site expansion. Resour Conserv Recycl 149:352–362

Ruktanonchai NW, DeLeenheer P, Tatem AJ, Alegana VA, Caughlin TT, zu Erbach-Schoenberg E, Lourenço C, Ruktanonchai CW, Smith DL (2016) Identifying malaria transmission foci for elimination using human mobility data. PLoS Comput Biol 12(4):e1004846

Ruktanonchai NW, Floyd JR, Lai S, Ruktanonchai CW, Sadilek A, Rente-Lourenco P, Ben X, Carioli A, Gwinn J, Steele JE, Prosper O (2020) Assessing the impact of coordinated COVID-19 exit strategies across Europe. Science 369:1465–1470. https://doi.org/10.1126/science.abc5096

Salat H, Smoreda Z, Schläpfer M (2020) A method to estimate population densities and electricity consumption from mobile phone data in developing countries. PLoS ONE 15(6):e0235224. https://doi.org/10.1371/journal.pone.0235224

Schmid T, Bruckschen F, Salvati N, Zbiranski T (2017) Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. J R Stat Soc Ser A Stat Soc 180(4):1163–1190. https://doi.org/10.1111/rssa.12305

Siła-Nowicka K, Vandrol J, Oshan T, Long JA, Demšar U, Fotheringham AS (2016) Analysis of human mobility patterns from GPS trajectories and contextual information. Int J Geogr Inf Sci 30(5):881–906

Sonop Farms (2020) Products. https://sonop.com.na/availability/. Accessed 14 Dec 2020

Steele JE, Sundsøy PR, Pezzulo C, Alegana VA, Bird TJ, Blumenstock J, Bjelland J, Engø-Monsen K, de Montjoye YA, Iqbal AM, Hadiuzzaman KN (2017) Mapping poverty using mobile phone and satellite data. J R Soc Interface 14(127):20160690

Steenbruggen J, Trano E, Nijkamp P (2015) Data from mobile phone operators: A tool for smarter cities? Telecommun Policy 39(3-4):335–346

Stevens FR, Gaughan AE, Linard C, Tatem AJ (2015) Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. PLoS ONE 10(2):e0107042. https://doi.org/10.1371/journal.pone.0107042

Sorichetta A, Hornby GM, Stevens FR, Gaughan AE, Linard C, Tatem AJ (2015) High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. Sci Data 2:150045. https://doi.org/10.1038/sdata.2015.45

Tang J, Liu F, Wang Y, Wang H (2015) Uncovering urban human mobility from large scale taxi GPS data. Physica A 438:140–153

Tatem AJ, Huang Z, Narib C, Kumar U, Kandula D, Pindolia DK, Smith DL, Cohen JM, Graupe B, Uusiku P, Lourenço C (2014) Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning. Malaria J 13(1):52

The Namibia Ministry of Health and Social Services and ICF International (2014) The Namibia demographic and health survey 2013. The Namibia Ministry of Health and Social Services, Windhoek, Namibia, and ICF International, Rockville, MD, USA. https://dhsprogram.com/pubs/pdf/FR298/FR298.pdf. Accessed 29 Mar 2022

The Namibian. National News (2011) Lack of housing angers Aussenkehr residents. https://web.archive.org/web/20120606180922/http://www.namibian.com.na/news-articles/national/full-story/archive/2011/october/article/lack-of-housing-angers-aussenkehr-residents/. Accessed 16 Nov 2020

UNICEF (2016) The State of the World's Children 2016 statistical tables. https://data.unicef.org/resources/state-worlds-children-2016-statistical-tables/. Accessed 16 Nov 2020

Vanhoof M, Lee C, Smoreda Z (2018) Performance and sensitivities of home detection from mobile phone data. In: Hill CA, Biemer PP, Buskirk TD, Japec L, Kirchner A, Kolenikov S, Lyberg L (eds) Big data meets survey science: a collection of innovative methods, 1st edn. Wiley

Vazquez-Prokopec GM, Bisanzio D, Stoddard ST, Paz-Soldan V, Morrison AC, Elder JP, Ramirez-Paredes J, Halsey ES, Kochel TJ, Scott TW, Kitron U (2013) Using GPS technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment. PLoS ONE 8(4):e58802

Wesolowski A, Eagle N, Noor AM, Snow RW, Buckee CO (2013) The impact of biases in mobile phone ownership on estimates of human mobility. J R Soc Interface 10(81):20120986

Xiao Y, Wang D, Fang J (2019) Exploring the disparities in park access through mobile phone data: evidence from Shanghai, China. Landsc Urban Plan 181:80–91

Xu Y, Shaw SL, Zhao Z, Yin L, Fang Z, Li Q (2015) Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach. Transportation, 1 42(4):625–46. https://doi.org/10.1007/s11116-015-9597-y

Xu Y, Belyi A, Bojic I, Ratti C (2017) How friends share urban space: An exploratory spatiotemporal analysis using mobile phone data. Trans GIS 21(3):468–487

Zhang G, Rui X, Poslad S, Song X, Fan Y, Ma Z (2019) Large-scale, fine-grained, spatial, and temporal analysis, and prediction of mobile phone users' distributions based upon a convolution long short-term model. Sensors 19(9):2156

Zhao Z, Shaw SL, Xu Y, Lu F, Chen J, Yin L (2016) Understanding the bias of call detail records in human mobility research. Int J Geogr Inf Sci 30(9):1738–1762

zu Erbach-Schoenberg E, Alegana VA, Sorichetta A, Linard C, Lourenço C, Ruktanonchai NW, Graupe B, Bird TJ, Pezzulo C, Wesolowski A, Tatem AJ (2016) Dynamic denominators: the impact of seasonally varying population numbers on disease incidence estimates. Popul Health Metrics 14(1):35

Zufiria PJ, Pastor-Escuredo D, Úbeda-Medina L, Hernandez-Medina MA, Barriales-Valbuena I, Morales AJ et al. (2018) Identifying seasonal mobility profiles from anonymized and aggregated mobile phone data. Application in food security. PLoS ONE 13(4):e0195714. https://doi.org/10.1371/journal.pone.0195714

Acknowledgements

The authors would like to thank Mobile Telecommunications Limited for providing access to the mobile phone data. AS, DW, AC, MB, PK, GER, JS are all supported by the Bill & Melinda Gates Foundation (Grant Number OPP1134076). SL is supported by funding from the Bill & Melinda Gates Foundation (OPP1134076, INV-024911), the EU H2020 (MOOD 874850), and the National Natural Science Foundation of China (81773498). AJT is supported by funding from the Bill & Melinda Gates Foundation (OPP1106427, OPP1032350, OPP1134076, OPP1094793), the Clinton Health Access Initiative, the UK Foreign, Commonwealth and Development Office (UK-FCDO), the Wellcome Trust (106866/Z/15/Z, 204613/Z/16/Z), the National Institutes of Health (R01AI160780), and the EU H2020 (MOOD 874850). CEU is supported by funding from the Bill and Melinda Gates Foundation and Gavi, the Vaccine Alliance.

Author contributions

AS and JS conceptualised the study. EZE-S and LS contributed to data acquisition and pre-processing. AS, JS, DW, AC developed the methodology. DW and AC wrote the original draft and performed the main analysis including distribution of CDRs, ranking, decomposition and multivariate clustering and produced tables and figures. AS, DW, AC reviewed and edited the manuscript. CEU provided guidance on the statistical analysis

for ranking and clustering. AJT, CWR, MB, CEU, JS and AS contributed to the interpretation of findings. AJT, JS and AS acquired funding. All authors have read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Ethical approval

Ethical clearance for collecting and using secondary data in this study was granted by the Institutional Review Board of the University of Southampton (No. 48002). All data were supplied and analysed in an anonymous format, without access to personal identifying information.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-022-01256-8.

Correspondence and requests for materials should be addressed to D. Woods.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

D. Woods[1]✉, A. Cunningham[1], C. E. Utazi[1], M. Bondarenko[1], L. Shengjie[1], G. E. Rogers[1], P. Koper[1], C. W. Ruktanonchai[1,2], E. zu Erbach-Schoenberg[1], A. J. Tatem[1], J. Steele[1] & A. Sorichetta[1]

[1]WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton SO17 1BJ, UK. [2]Population Health Sciences, College of Veterinary Medicine, Virginia Tech, Blacksburg, VA, USA. ✉email: dw@geodata.soton.ac.uk