ARTICLE

Exploring the use of mobile phone data for national migration statistics

Shengjie Lai [1,2,3], Elisabeth zu Erbach-Schoenberg[1,2], Carla Pezzulo[1], Nick W. Ruktanonchai[1,2], Alessandro Sorichetta[1,2], Jessica Steele[1], Tracey Li[2], Claire A. Dooley[1,2] & Andrew J. Tatem[1,2]

ABSTRACT   Statistics on internal migration are important for keeping estimates of sub-national population numbers up-to-date, as well as urban planning, infrastructure development, and impact assessment, among other applications. However, migration flow statistics typically remain constrained by the logistics of infrequent censuses or surveys. The penetration rate of mobile phones is now high across the globe with rapid recent increases in ownership in low-income countries. Analyzing the changing spatiotemporal distribution of mobile phone users through anonymized call detail records (CDRs) offers the possibility to measure migration at multiple temporal and spatial scales. Based on a dataset of 72 billion anonymized CDRs in Namibia from October 2010 to April 2014, we explore how internal migration estimates can be derived and modeled from CDRs at subnational and annual scales, and how precision and accuracy of these estimates compare to census-derived migration statistics. We also demonstrate the use of CDRs to assess how migration patterns change over time, with a finer temporal resolution compared with censuses. Moreover, we show how gravity-type spatial interaction models built using CDRs can accurately capture migration flows. The results highlight that estimates of migration flows made using mobile phone data is a promising avenue for complementing more traditional national migration statistics and obtaining more timely and local data.

[1] WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton SO17 1BJ, UK. [2] Flowminder Foundation, SE-113 55 Stockholm, Sweden. [3] School of Public Health, Key Laboratory of Public Health Safety of the Ministry of Education, Fudan University, 130 Dongan Road, 200032 Shanghai, China. These authors contributed equally: Shengjie Lai, Elisabeth zu Erbach-Schoenberg   Correspondence and requests for materials should be addressed to S.L. (email: laishengjie@foxmail.com) or to A.J.T. (email: Andy.tatem@gmail.com)

Introduction

Human populations are highly mobile in the modern world, and migration is one of the main factors that determines changes in population size, distribution, and structure (Abel and Sander, 2014; Agliari et al. 2018). As migration impacts the demographic and socio-economic aspects of a country, it has become one of the most challenging issues confronting policymakers for nations around the world (International Organization for Migration, 2017a, c). Understanding internal migration, which is normally substantially larger than international migration rates, and their changes over time is critical for keeping subnational population numbers up-to-date (Frayne, 2005; Pendleton et al. 2014; Wardrop et al. 2018). Contemporary data on internal migration flows are valuable for urban planning, resource allocation, infrastructure development, public service provision, and impact assessments. For instance, identifying where people migrate internally is often vital in development work, as migrants might be marginalized and at higher risk due to a lack of resources to meet demands (Lu et al. 2012; Lu et al. 2016; Ruktanonchai et al. 2016a). However, our knowledge of contemporary internal migration patterns remains poor for many countries (Garcia et al. 2015; Sorichetta et al. 2016; International Organization for Migration, 2017b), and is difficult to update between data collections for the majority of countries around the world.

Data collected from traditional sources, such as national population and housing censuses and household surveys, are the primary source for migration statistics (International Organization for Migration, 2018). Within population and housing censuses, migration is typically measured through a change in residence over a 1-year or 5-year period prior to the census. The increasing use of global positioning systems (GPS) has supported the collection of more spatially precise data, but each census only provides a single snapshot of migration flows, commonly once every decade, and migration patterns typically change over time between censuses or surveys (Namibia Statistics Agency, 2013; Wesolowski et al. 2013). Moreover, surveys only sample a small proportion of population, and the logistical challenge of censuses makes them an infrequent and expensive source of demographic data (Wardrop et al. 2018).

Moreover, as migration is anticipated to continue to rise, both in terms of volume and reach, the need for timely updates to demographic statistics and inform migration policy development increases–a need that traditional sources are typically not well-equipped to meet (International Organization for Migration, 2018). To predict contemporary migration for many countries, a growing interest in the modeling of migration flows emerged, leading to the advanced development of modeling methodologies to estimate migration rates (Courgeau, 1995; Henry et al. 2003; Cohen et al. 2008; Abel, 2013; Abel and Sander, 2014; Garcia et al. 2015; Sorichetta et al. 2016; Vobruba et al. 2016). However, regardless of how sophisticated these methods are, these estimates remain largely constrained by the lack of contemporary input data and often their coarse spatiotemporal resolution (Garcia et al. 2015; Sorichetta et al. 2016).

Call detail records (CDRs) routinely collected by mobile phone operators for billing purposes are particularly promising for analyzing migration-related phenomena and a potential solution to existing data gaps (International Organization for Migration, 2018). CDRs contain an entry for each call or text (or other billable event) made or received by any anonymous user, together with the date and time of each communication and an identifier for the tower that the communication was routed through within the operator's network (Ruktanonchai et al. 2016b; Zu Erbach-Schoenberg et al. 2016). Then the tower-level location of each communication can be identified, and from this, spatially and temporally explicit estimates of human mobility, which can be derived from anonymised CDRs from the movement of individual mobile user between different communications. These data have been increasingly used for quantifying short-term human mobility, mapping dynamically changing population densities, estimating infectious disease spread risk, and measuring population displacements due to disasters and conflicts (Lu et al. 2012; Wesolowski et al. 2012; Deville et al. 2014; Tatem et al. 2014; Wesolowski et al. 2014a; Wesolowski et al. 2015a; Wesolowski et al. 2015c; Lu et al. 2016; Ruktanonchai et al. 2016b; Zu Erbach-Schoenberg et al. 2016; Wesolowski et al. 2017). Moreover, previous work on defining overall and seasonal patterns of population movement using CDRs suggested they could also be used to model internal migration (Blumenstock, 2012; Wesolowski et al. 2013; Ruktanonchai et al. 2016a; Wesolowski et al. 2017).

In previous studies, however, CDRs frequently spanned much shorter periods than one year, or multi-year mobility analysis using CDRs have been presented, but no studies have compared individual places of usual residence across different years to estimate migration flows by matching the definition of migration used in censuses (Blumenstock, 2012; Zu Erbach-Schoenberg et al. 2016; Wesolowski et al. 2017). Based on a multiannual CDR dataset in Namibia, for the first time, we assess how CDRs as a novel data source might be used efficiently and accurately to replicate the internal migration statistics produced in a census, and examine how CDRs could improve the estimates made using classical gravity models. This study also reveals otherwise unmeasurable year-by-year migration patterns to assess the potential of CDRs for updating internal migration statistics.

Datasets

Census migration statistics. The most recent census in Namibia was conducted in 2011, and we obtained the internal migration statistics between regions from a census-based migration report published by the Namibia Statistics Agency in 2015 (Namibia Statistics Agency, 2015). To derive the 1-year internal migration statistics, the census (with a reference night of 28 August 2011) asked about each individual's place of usual residence (where does the person usually live?) and the place of previous residence (where did the person usually live since September 2010?). The place of residence refers to the location where a person usually lives for the majority part of any year (at least six months). An individual was considered as an internal migrant if the regions of usual residence and previous residence did not match in the 2011 census.

CDR-derived flow data. To assess whether mobile network data could produce comparable migration statistics, we obtained a large dataset of anonymized 72 billion CDRs between October 2010 and April 2014 from Mobile Telecommunications Limited (MTC) (Mobile Telecommunications, 2018) (Mobile Telecommunications, 2018) (Mobile Telecommunications, 2018). MTC is the leading network operator in Namibia with a 76% market share and providing network spatial coverage 95% population (Mobile Telecommunications, 2018). The CDR dataset obtained from MTC included the time and routing tower for each call and text and a random uniquely hashed number for each user. The approximate location of a user was defined by the location of the routing mobile phone tower for each communication. The data were spatially aggregated to regional level to match the census migration data and to further reduce sensitivities of using individual level data. We estimated a user's place of residence for a given period as the region where the user was observed most frequently during the period of interest. As the

data on very infrequent mobile phone users or seasonal movement (e.g., short-term travels in holidays), might introduce noise in defining residential places, we only included any user who was active for more than 30 days each year (12 months) defined as below.

To match as closely as possible the time frame used in census and to be comparable between the 2011 and 2012 periods, we defined the residence of each user for each year: Year 1 (October 2010–September 2011), Year 2 (October 2011–September 2012), and Year 3 (October 2012–September 2013), respectively. We derived migration flows of mobile phone users for periods 2011 and 2012 by comparing residences between Years 1 and 2, and between Years 2 and 3, respectively. If mobile users changed residence between the two years, they were identified as migrants, otherwise as non-migrants. In addition, we also assessed the potential impact of data filtering and different time lengths on defining residences (Supplementary information [SI] text).

Model covariates. For estimating migration by models for the 2011 period, we also collated potential migration-related demographic, socioeconomic, geographic, and environmental variables, as described in previous studies (Garcia et al. 2015; Sorichetta et al. 2016), including population by region in 2010 and 2011 (Namibia Statistics Agency, 2013); the proportions of population living in urban areas, male population, population aged 15–59, educated population, labor force participation, and marital status in population at aged 15 years and above; administrative unit boundaries to define the distance and contiguity between regions and their area (Zhao et al. 2012); and the average annual precipitation by region. The collation of covariates is detailed in the SI Text.

Models and analysis

We fit three types of models to census data to explore whether CDR-derived migration data can accurately replicate traditional census-derived migration statistics. Three types of models were included (Table S1): (1) CDR-based linear models (CDRLMs), simply using CDR-derived migrating user data alone or combined with covariates used in gravity models; (2) gravity-type spatial interaction models (GTSIMs), which have been applied extensively to estimate migration flows based on a range of migration-related push-pull factors including populations and distance between origin and destination (Zipf, 1946; Hua and Porell, 1979; Garcia et al. 2015; Wesolowski et al. 2015b; Ruktanonchai et al. 2016a; Sorichetta et al. 2016; Vobruba et al. 2016); and (3) GTSIMs extended using CDR data (thereafter called CGTSIMs).

CDR-based linear models. Initially, we used *Pearson* correlation coefficients to assess the relationship between CDR and census data. To investigate how well the CDRs can replicate the census migration numbers, we built four sub-models of CDRLMs using independent variables of CDR-derived migrating user numbers or integrating with other covariates:

$$\mathrm{MIG}_{i,j} = \beta_0 + \beta_1 \mathrm{CDR}_{i,j} + \overrightarrow{\beta}[X] \tag{1}$$

where the dependent variable $\mathrm{MIG}_{i,j}$ is comprised of the observed migration flows between regions in Namibia from the census. $\mathrm{CDR}_{i,j}$ is the number of CDR-derived migrations from origin i to destination j, with the coefficient β_1 and the constant β_0. The suite of models was built by successively adding same covariates that were used in GTSIMs and represented by the matrix X and its vector of coefficients $\overrightarrow{\beta}$.

Gravity-type spatial interaction models. In the simplest form of gravity models (Zipf, 1946), the flow of migration between regions is proportional to their total populations and inversely proportional to the distance between them:

$$\mathrm{MIG}_{i,j} = \frac{\mathrm{POP}_i^{\beta_1}\mathrm{POP}_j^{\beta_2}}{\mathrm{DIST}_{i,j}^{\beta_3}} \tag{2}$$

where POP_i and POP_j refer to populations at an origin i and a destination j in 2010, respectively; $\mathrm{DIST}_{i,j}$ represents the distance between i and j; The exponents, β_1, β_2, and β_3, are used to indicate the magnitude of the effect for each variable.

As a range of potential push-pull factors, e.g., urbanization and natural disaster, could affect human migration, the models can be further extended to reach more accurate estimates as described in previous studies (Garcia et al. 2015; Sorichetta et al. 2016). However, given that the number of regions in Namibia is small (13 regions) and to prevent overfitting, we only tested models by replacing the total population variables with the percentage of population living in urban areas (URBAN_i and URBAN_j) and the precipitation (RAIN_i and RAIN_j) in origin and destination, respectively (SI text). Although both logistic and *Poisson* regressions have been widely used in gravity models to predict migration flows, the outputs from logistic regression should be identical to estimates of *Poisson* regression by adding an offset variable of non-migrating populations (Garcia et al. 2015; Ruktanonchai et al. 2016a; Sorichetta et al. 2016). Therefore, we only fit GTSIMs using the logistic regression function here:

$$\frac{\mathrm{MIG}_{i,j}}{\mathrm{TOT}_i} = \frac{e^{\beta_0+\beta_1 P_i+\beta_2 P_j-\beta_3 \mathrm{DIST}_{i,j}}}{1+e^{\beta_0+\beta_1 P_i+\beta_2 P_j-\beta_3 \mathrm{DIST}_{i,j}}} \tag{3}$$

where TOT_i represents the total population residing in an origin i in 2010, and where P_i and P_j refer to the push factor at origin and pull factor at destination, respectively (Table S1). Moreover, the CGTSIMs with additional CDRs variables were tested to assess how well the CDR-derived migration data could improve the performance of gravity models.

Model comparisons. By fitting to census statistics for each model, we used a leave-one-out-cross-validation approach (Hastie et al. 2009) to split the dataset to calculate the goodness-of-fit indicators, including root-mean-square error (RMSE), R-squared (R^2) and Akaike Information Criterion (AIC). The model with the lowest RMSE was determined as the best model of each model family. The estimates of migration between regions were then calculated using the optimal model, and the inflow, outflow and netflow for each region in Namibia were also aggregated.

As our models used non-spatial regression approaches, and spatial autocorrelation may exist in migration data (Tobler, 1970; Getis, 2008; Sorichetta et al. 2016), a shuffle test was used to assess whether any spatial dependencies significantly affected the performance of our models. First, we randomly permuted the census-derived migration data across all regions. Then each model was fitted to calculate RMSE by using each shuffled dependent variable, and the distribution of RMSE could be produced through 1000 iterations. If the "real" RMSE of each model that was fitted with the "ground truth" migration data was less than all 1000 simulated values of RMSE using the shuffled data, we assumed that the spatial dependencies were not significant in our models. All analyses were done within the *R* statistical environment (version 3.5.2), and fitting procedures of models were conducted using *caret* Package (Kuhn, 2008; R Core Team, 2018).

Estimating migration for the 2012 period. Due to the lack of migration statistics in 2012 for fitting models in Period 2012, the CDRLM using only CDR data and its coefficients fitted for Period 2011 were used to predict the migration for Period 2012 and compare the pattern of migration across periods. Moreover, to account for increasing numbers of mobile phone users from 2011, the CDR-derived data for migrating users for Period 2012 were inversely weighted by the increasing rate of mobile phone users for each region to offset the potential bias introduced by increasing mobile ownership across periods.

Mobile phone ownership analysis. As mobile phone users only represent a proportion of the whole population, we utilized data from the 2013 Namibia DHS (The Namibia Ministry of Health et al. 2014) to assess the extent to which there is a possible exclusion of certain groups at a household level within the CDRs in the context of Namibia (SI text). To account for potential mobile phone ownership biases across regions, the models mentioned above were also tested by using CDR data adjusted by two approaches respectively: (1) using the proportion of mobile phone ownership to inversely weight CDR-derived migration data by region; and (2) adding the proportion of ownership as an additional variable into models.

Results

Correlations between census-derived and CDR-derived migrations. According to the 2011 Namibia population and housing census (Namibia Statistics Agency, 2015), a total of 40,867 Namibian (2.0% of 2,013,671 people) migrated by changing their places of residence between regions in Namibia over the one-year period prior to the census in August 2011, with the highest migration into Khomas, the capital region of Namibia, and the highest migration out from the Zambezi region in the northeast of Namibia (see Fig. 1 and S1). Based on the anonymized CDRs in Namibia between October 2010 and April 2014,

we estimated the number of migrating mobile users by comparing their residences between two years of October 2010–September 2011 and October 2011–September 2012 in Period 2011 (SI text; Figs S2 and S3). A high correlation (Pearson's coefficient, $r = 0.91$) was found between the numbers of census-derived population and mobile phone users included in Period 2011 (Fig. 2a). Furthermore, the migration flows were also highly correlated ($r = 0.84$) between census data and CDR-derived 117,173 migrating mobile users (11.2% of 1,049,379 users) (Figs S4 and S5).

Substantial differences in the Zambezi region were observed when comparing the census and CDR data, with more census-derived migrants than from the CDRs (Figs S5 and S6). The Zambezi region lost a significant proportion of its population (5.5%), which was attributed to displacement due to floods in the period of April-June 2010, out of the time frame of the census and CDRs (International Federation of Red Cross and Red Crescent Societies, 2011; Namibia Statistics Agency, 2015). According to definitions used in census (SI text) (Namibia Statistics Agency, 2015), if people moved to the places of displacement before September 2010 and still lived in the same places by the time of census, they should be considered as non-migrants. Therefore, the displaced populations from Zambezi before September 2010 may well have been misclassified as migrants in the census. Moreover, based on the data of CDR-derived monthly residence, the inflow and outflow of Zambezi seem to be seasonal without aberrational high movements from October 2010 to April 2014 (Fig. S7). After removing the data from Zambezi, the relationships between census-derived and CDR-derived migration data significantly improved, with the r value increasing from 0.84 to 0.96. Therefore, we present the following results without the Zambezi region, and relevant comparable analyses for all regions are provided in the SI.

Comparing migration prediction models. In general, the goodness-of-fit indicators, including RMSE, R^2, and AIC, show
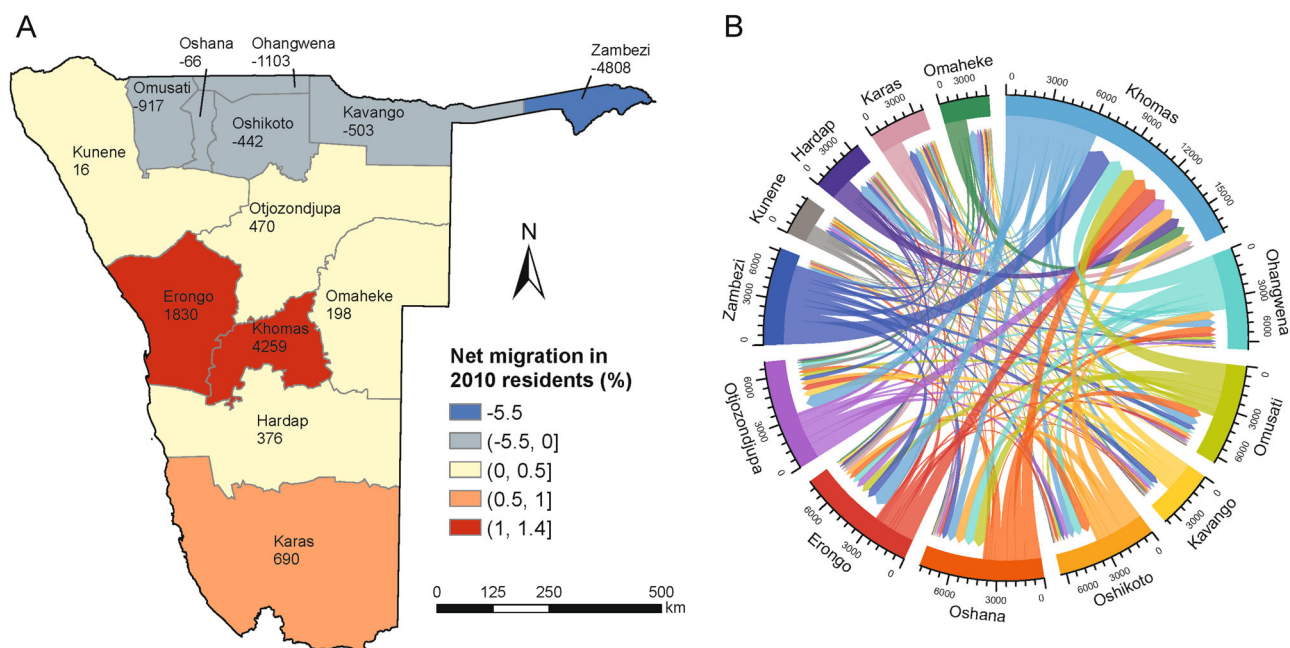

Fig. 1 Census-derived internal migration in Namibia, September 2010–August 2011. **a** Net migration by region. The number of net migrants by region is presented under the name of each region. **b** Circular plot of migrant flows between regions. The origins and destinations of migrants are each assigned a color and represented by the circle's segments. The direction of the flow is encoded by both the origin region's color and a gap between the flow and the destination region's segment. The volume of movement is indicated by the width of the flow at the beginning and end points. Tick marks on the circle segments show the number of migrants (inflows and outflows)
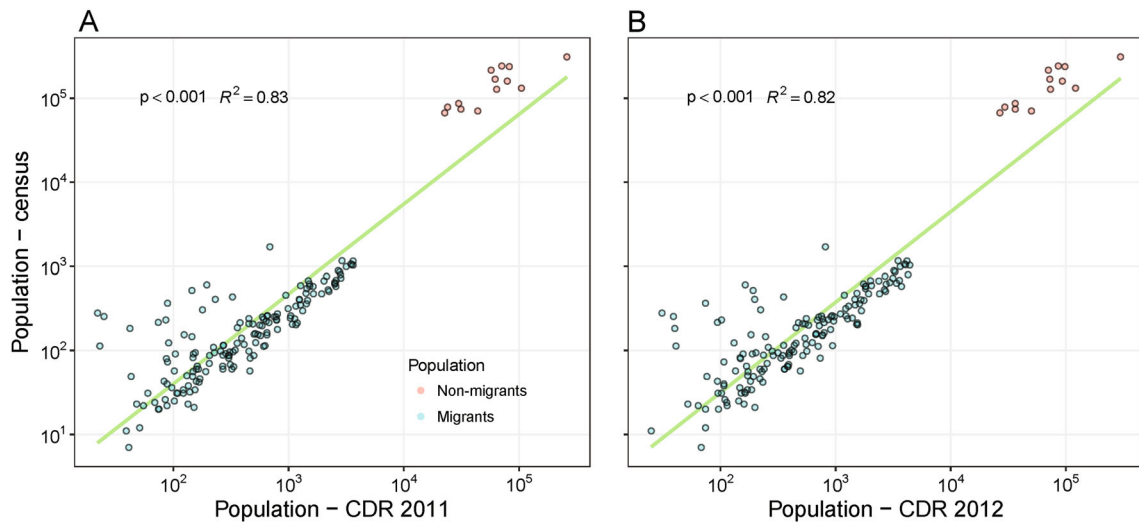
Fig. 2 Logarithmic relation between census-derived populations in 2011 and the number of CDR-derived mobile phone users in Periods 2011 (**a**) and 2012 (**b**) at regional level. The green solid lines represent linear regression fit, with p and R^2 values provided
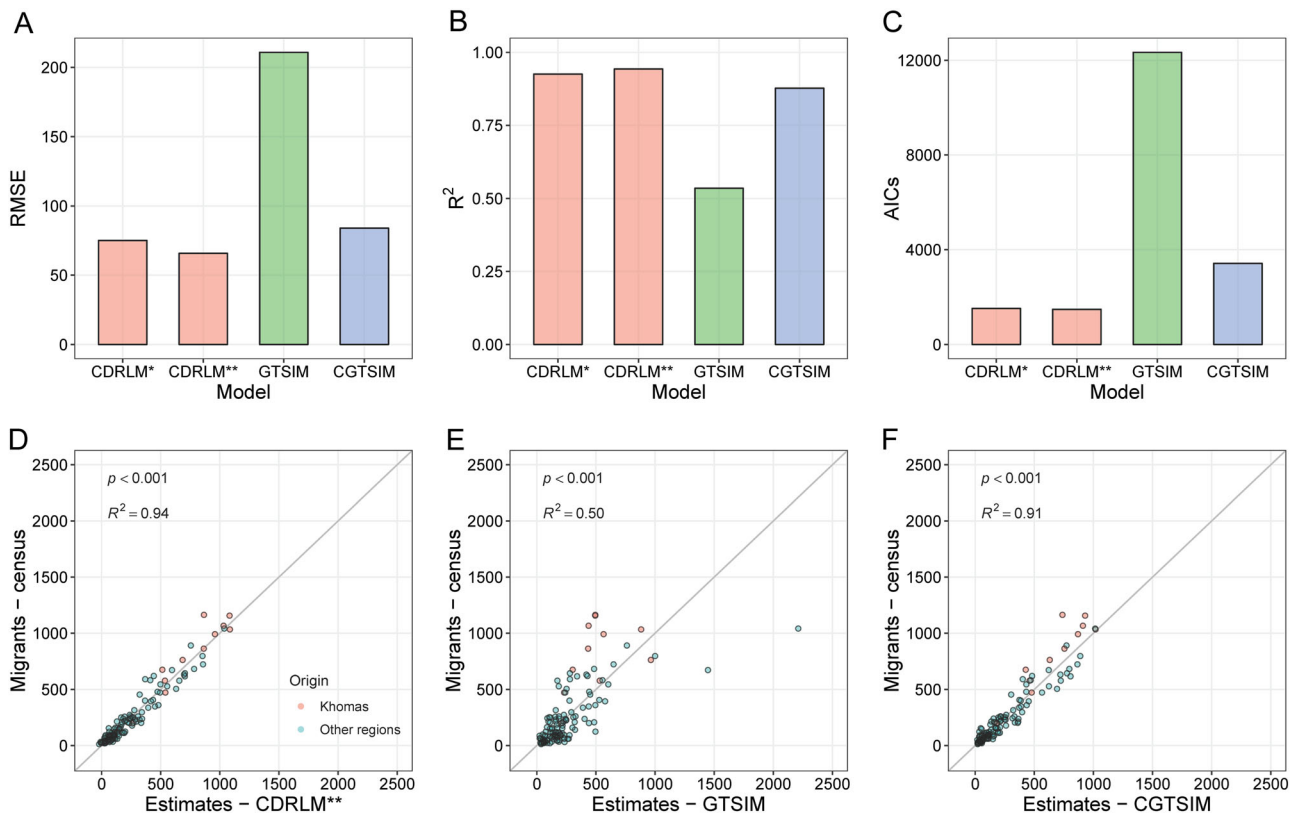

Fig. 3 Precision and accuracy assessments of models for replicating 2011 census-derived migration statistics. The indicators of **a** root-mean-square error (RMSE), **b** R^2, and **c** Akaike Information Criterion (AIC), were computed to compare three types of models: CDR-based linear model (CDRLM), gravity-type spatial interaction model (GTSIM), and CDR-based GTSIM (CGTSIM). The scatterplots of census data versus estimates using models are presented in **d**–**f**, respectively. The Zambezi region as an outliner is excluded, and unadjusted CDR data are used. For GTSIM and CGTSIM, only models with the lowest RMSE are showed here. The formula and results of all models are presented in Table S1 and Figs S8–S10. *The model #1 of CDRLMs. **The model #3 of CDRLMs

that CDRLMs using only CDR data could precisely and accurately replicate census-derived statistics, with a better predictability than GTSIMs (Figs S8–S10). Moreover, the performance of GTSIMs could be substantially improved by using CDRs. Comparing the "real" RMSE with the distributions of RMSEs generated by the shuffled census data, it was evident that spatial autocorrelation was not significant in our models (Fig. S11). According to the optimized model with the lowest RMSE, all three families of models could capture the patterns of migration flows between regions (Fig. S12), but CDRLMs had a higher accuracy in predictions compared with GTSIMs and CGTSIMs (Fig. 3). Additionally, in terms of outflow, inflow, and net
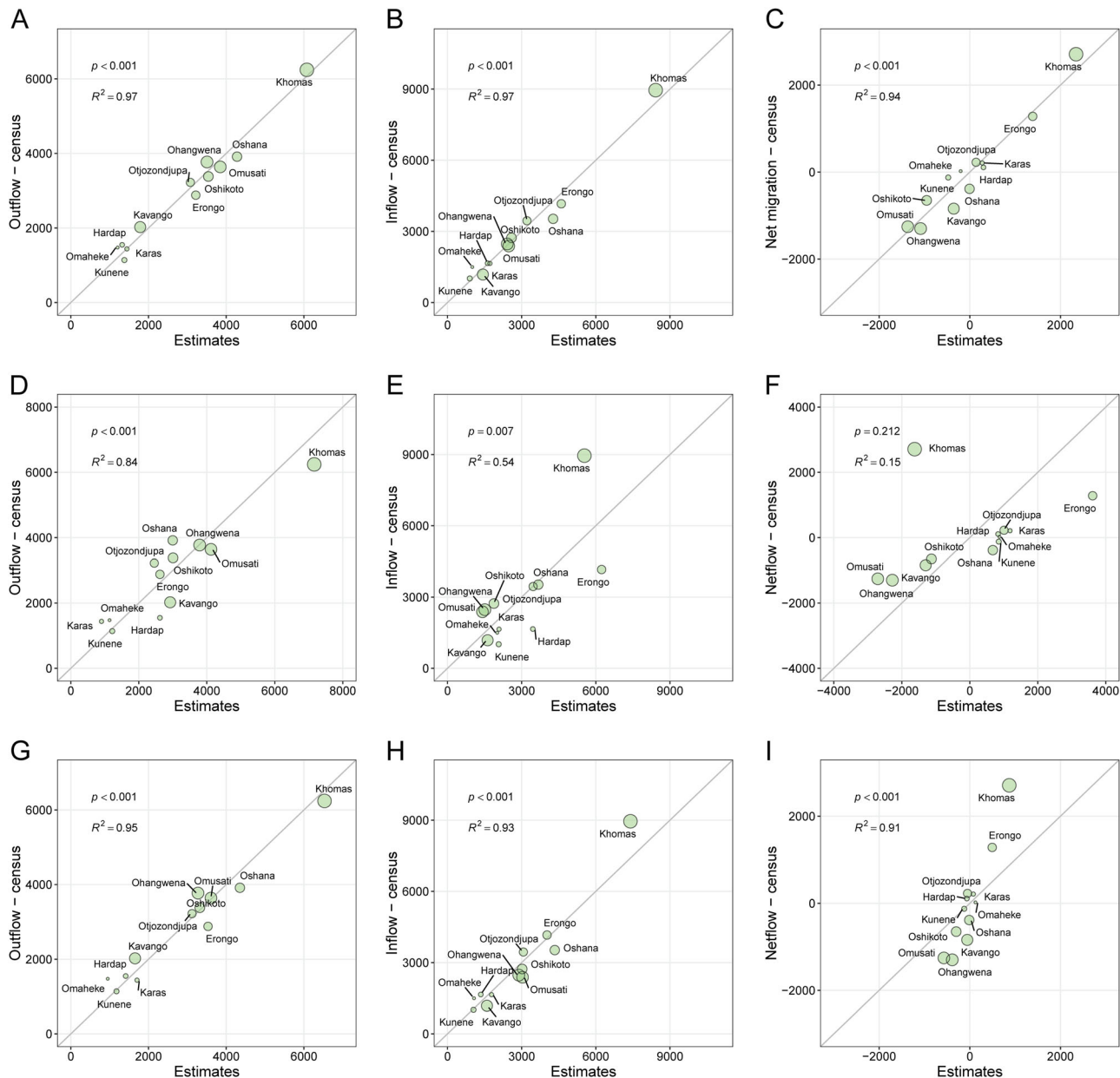
Fig. 4 Comparing regional outflow, inflow and net migration between census data and models' estimates for Period 2011: CDRLM in **a**–**c**, GTSIM in **d**–**f**, and CGTSIM in **g**–**i**. For each type of model, the results of model with the lowest RMSE are presented. The Zambezi region as an outliner is excluded, and unadjusted CDR data are used

migration aggregated by region, the estimates from CDRLM were highly correlated ($R^2 = 0.97$, 0.97, and 0.94 respectively) with the census-derived data (Fig. 4 and S13).

Mobile phone ownership bias and model adjustment. As mobile users only represent a proportion of the population, to understand the potential phone ownership bias, we utilized data from the 2013 Namibia Demographic and Health Survey (DHS) (The Namibia Ministry of Health et al. 2014) to assess to the extent to which there is a possible exclusion of certain groups with specific characteristics from CDRs in Namibia. The 2013 DHS reported that although the large majority (88.5%) of households interviewed owned at least one mobile phone (The Namibia Ministry of Health et al. 2014), the lower-income and rural households with older and uneducated heads were less likely to be able to afford a cell phone, and there was a significant

ownership differential between regions in Namibia (SI text; Tables S2 and S3). To account for the potential mobile ownership bias between regions, two approaches were used to adjust CDRs, respectively. However, the performance of both CDRLMs and CGTSIMs were not significantly improved by these adjustments (Figs S8–S10).

Predicting migration in 2012. The multiannual time series of CDRs in Namibia allows us to assess their potential to be used to update intercensal national statistics and understand the changing patterns of internal migrations across years. By comparing the places of residence between the two years of October 2011–September 2012 and October 2012–September 2013 (hereafter called Period 2012), we captured 144,064 migrants in 1,238,124 mobile users, with a similar proportion of 11.6% as Period 2011. The increasing numbers of migrations between
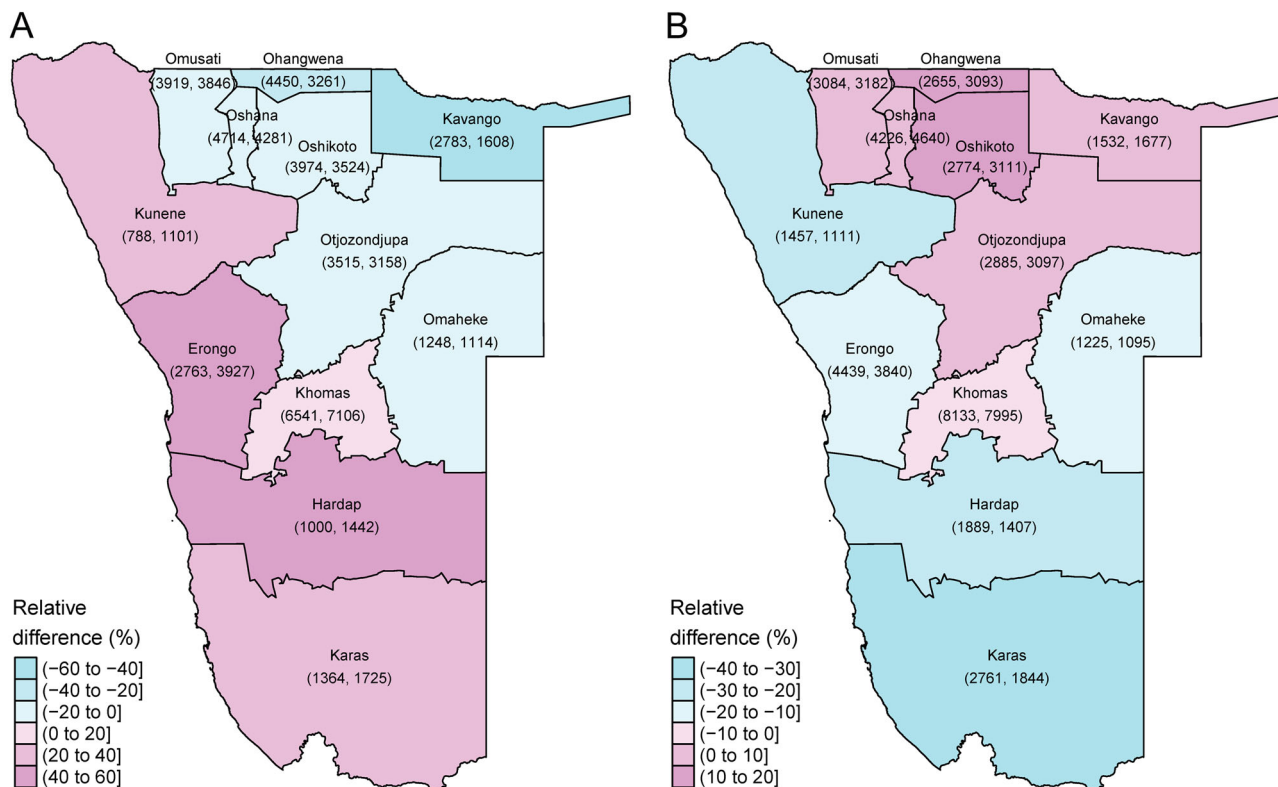
Fig. 5 Relative difference of regional outflow **a** and inflow **b** between Period 2011 and Period 2012. The migrations were estimated by the CDRLM using only CDRs, and the adjusted CDR data of Period 2012 were used to offset the impact of the increasing mobile phone ownership across periods. The numbers of migrants by region in Period 2011 and Period 2012 are presented under the name of each region, respectively, and the Zambezi region as a significant outliner is excluded

periods was likely due to the increasing penetration rate of mobile phones across years (Figs S4 and S14). To compare migration patterns between two periods, we adjusted the number of CDR-derived migrating users in Period 2012 by region to offset the increasing mobile phone ownership across periods. Then, the simplest CDRLM using only CDR data and its coefficients estimated for Period 2011 were used to predict migration for Period 2012 using the corresponding adjusted CDR data (Fig. S14). We observed highly consistent patterns of migration flows between Periods 2011 and 2012 as well as the outflows, inflows and net migration aggregated by region (Figs S14–S16). However, the relative differences across periods show greater variations in outflow than in inflow between regions, with more people moving out from the West-South regions and into the northern regions in Namibia (see Fig. 5).

Discussion

Migration is difficult to measure frequently, particularly at local scales, and data from censuses are typically collected just once every decade, pushing a need for innovation in the production of migration statistics (International Organization for Migration, 2018). The penetration rate of mobile phones is now high across the globe, and analyzing the changing spatiotemporal distribution of mobile phone users through anonymized CDRs offers the possibility to measure migration at multiple temporal and spatial scales. Global mobile phone network subscriber numbers passed the five billion mark in 2017 with a global penetration rate of 66%, and the number is forecasted to continue to grow, moving upto 71% by 2025, with rapid recent increases in ownership in low-income countries (The GSM Association, 2018). The data collected every second by mobile network operators have the potential to contribute to the "big data revolution" in complementing more

traditional statistics through updating internal migration statistics in a timely, accurate and low-cost way.

This study demonstrates how the analysis of CDRs can replicate national internal migration statistics to complement outputs from censuses. The multiannual time series of CDRs with high spatiotemporal resolution facilitates the derivation of residence measures, matching closely the definitions used in censuses. We found that not only can the estimates of migration produced through CDRs be as accurate as census data-derived measures, but these data offer additional benefits in terms of updating intercensal migration numbers and understanding changing patterns of annual internal migration. Additionally, the methodologies presented are designed to be easy to implement while considering the impact of heterogeneous phone ownership across regions and years, and the simple linear model built using CDRs results in estimates with high precision and accuracy.

Results here suggest that CDRs can also improve the performance of gravity models. The GTSIMs explicitly state the spatial interaction relationship between migration and the push-pull factors that represent the benefits and costs of migration (Zipf, 1946; Hua and Porell, 1979). The estimates made using gravity models contribute to a better understanding of migration patterns, with known boundaries to their accuracy in the absence of censuses or surveys. However, due to the lack of high spatiotemporal resolution input data on contemporary population movements, such models used in previous studies resulted in high uncertainties in estimates (Garcia et al. 2015; Sorichetta et al. 2016; Vobruba et al. 2016). Though biases exist, as CDR-derived migration data directly relate to populations who moved across the country over years, a combination of CDRs and other migration-related covariates could facilitate a significant improvement in the precision and accuracy of outputs from gravity models.

Internal migration is common in Namibia, and we estimated a larger number of migrating mobile phone users compared with those migrating within the census data. One reason is that CDRs do not suffer from recall bias (Wesolowski et al. 2014b) and capture missing data from people who moved, but did not register their previous residence in the census. Moreover, different time windows for data capture may also have contributed, with the CDR-based home definition window used here being wider than the census collection date. As elsewhere, the largest proportion of migration in Namibia is rural-to-urban migration, a phenomenon that relates partly to rapid urbanization (Garcia et al. 2015; Namibia Statistics Agency, 2015; International Organization for Migration, 2016). However, to accurately derive these migration flows and patterns using CDRs, any impacts from seasonal temporary movement should be minimized, such as holiday-related travel in December, patterns that are highly repetitive in Namibia (Zu Erbach-Schoenberg et al. 2016; Wesolowski et al. 2017). Using a 12-month time frame to define residence of mobile users may prevent bias of residence towards the temporary locations of seasonal travel (SI text). Further, the high temporal resolution of longitudinal CDRs enables the derivation and update of different statistical indicators of migration using varying periods, e.g., 2-year or 3-year migrations.

Some limitations must be acknowledged. First, to prevent overfitting and multicollinearity, our models did not test a large number of demographic, socioeconomic, geographic, and environmental factors and their combinations that might potentially affect migration as described before (Henry et al. 2003; Henry et al. 2004; Garcia et al. 2015; Wesolowski et al. 2015b; Ruktanonchai et al. 2016a; Sorichetta et al. 2016; Vobruba et al. 2016). Another methodological shortcoming is the lack of correction for spatial autocorrelation in the modeling by using a spatial regression model. However, a shuffle approach showed that any spatial dependencies likely did not significantly affect the performance of our models.

Mobile users only cover a proportion of the population, therefore, CDRs may provide an incomplete picture, not accounting for those who do not own and use a phone, mobile phone sharing, network coverage, or alternative networks. The spatiotemporal and demographic variations in the behavior of phone users can also bias population distribution and migration estimates (Lu et al. 2012; Deville et al. 2014). Mobile phone ownership typically biases toward more educated, urban males (SI text), and mobile network coverage may be substantially lower in remote rural locations (Wesolowski et al. 2017). However, a high proportion of the population in Namibia were SIM card owners that appeared in the CDRs (Stork, 2011), and a high share of ownership at household level was also found in the 2013 DHS data (The Namibia Ministry of Health et al. 2014). With continuously increasing mobile coverage and declining costs for handsets and network usage, the proportion of people owning and using mobile phones has been steadily increasing (The GSM Association, 2018), which will also decrease the influence of the problem of phone sharing, which is common in areas with low cell phone penetration.

In addition, to account for the impact of increasing user numbers across years on migration estimates, we adjusted the CDR-derived data for comparing interannual migration patterns, but these only represent an initial step for adjusting for mobile phone usage changes. Future studies on estimating migration could use other appropriate data, such as travel history and mobile phone use surveys to infer possible correlation in mobile use and migration in demographic-specific subgroups. In addition, due to the availability of data, we only investigated here internal migration over the course of a year. Long-term internal migration (>5 years) could be estimated by analyzing CDRs over a longer period and these could be integrated with additional data sources, such as Google Location History data (Ruktanonchai et al. 2018), to address relevant underlying research questions and technical issues in the future.

The results here show that estimates of migration flows made using CDRs is a promising avenue for complementing more traditional national statistics and obtaining more timely and local data. The metrics and approaches can inform distinctly different policy-relevant needs that require migration statistics and the implementation of policies geared towards providing relevant public services. Partnerships between governments and phone companies supported by appropriate incentives could enable accurate and rapid production of national migration statistics to complement census and survey-based data collection.

Data availability

The internal migration statistics between regions in Namibia in 2011 are available in the migration report published by the Namibia Statistics Agency in 2015 (https://cms.my.na/assets/documents/Migration_Report.pdf). The data of demographic and socioeconomic covariates used in this study were obtained from the main report of the Namibia 2011 Population and Housing Census (https://cms.my.na/assets/documents/p19dmn58guram30ttun89rdrp1.pdf). The administrative unit boundary at regional level matching the year of the census in Namibia is available at the Global Administrative Areas Database (https://gadm.org/maps/NAM_1.html), and the precipitation data can be obtained from the WorldClim version 2 (http://worldclim.org/version2). The call detail records datasets analyzed during the current study are 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.

References

Abel GJ (2013) Estimating global migration flow tables using place of birth data. Demogr Res 28:505–546

Abel GJ, Sander N (2014) Quantifying global international migration flows. Science 343(6178):1520–2

Agliari E, Barra A, Contucci P, Pizzoferrato A, Vernia C (2018) Social interaction effects on immigrant integration. Pal Commun 4:55

Blumenstock JE (2012) Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda. Inf Technol Dev 18(2):107–125

Cohen JE, Roig M, Reuman DC, GoGwilt C (2008) International migration beyond gravity: a statistical model for use in population projections. Proc Natl Acad Sci USA 105(40):15269–74

Courgeau D (1995) Migration theories and behavioural models. Int J Popul Geogr 1(1):19–27

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

Frayne B (2005) Survival of the poorest: migration and food security in Namibia. In: Mougeot Luc J A (ed) Agropolis, 1st edn. Taylor & Francis Group, London, 304 pages, pp. 31–44. https://doi.org/10.4324/9781849775892

Garcia AJ, Pindolia DK, Lopiano KK, Tatem AJ (2015) Modeling internal migration flows in sub-Saharan Africa using census microdata. Migr Stud 3(1):89–110

Getis A (2008) A history of the concept of spatial autocorrelation: a geographer's perspective. Geogr Anal 40(3):297–309

Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York

Henry S, Boyle P, Lambin EF (2003) Modelling inter-provincial migration in Burkina Faso, West Africa: the role of socio-demographic and environmental factors. Appl Geogr 23(2-3):115–136

Henry S, Piche V, Ouedraogo D, Lambin EF (2004) Descriptive analysis of the individual migratory pathways according to environmental typologies. Popul Environ 25(5):397–422

Hua C-i, Porell F (1979) A critical review of the development of the gravity model. Int Reg Sci Rev 4(2):97–126

International Federation of Red Cross and Red Crescent Societies (2011) DREF operation final report: Namibia Floods. https://reliefweb.int/sites/reliefweb.int/files/resources/DCE8D6D925CDFA2AC125782A00360D43-Full_Report.pdf. Accessed 25 April 2018

International Organization for Migration (2016) Migration in Namibia-a country profile 2015. https://cms.my.na/assets/documents/Migration_In_Namibia_-_Acountry_Profile_2015.pdf. Accessed 25 April 2018

International Organization for Migration (2017a) Data Bulletin-Global Migration Trends. https://publications.iom.int/system/files/pdf/global_migration_trends_data_bulletin_issue_1.pdf. Accessed 23 June 2018

International Organization for Migration (2017b) Data Bulletin - More than numbers: the value of migration data. https://publications.iom.int/system/files/pdf/global_migration_trends_capturing_value_issue_2.pdf. Accessed 23 June 2018

International Organization for Migration (2017c) World Migration Report 2018. https://publications.iom.int/system/files/pdf/wmr_2018_en.pdf. Accessed 25 June 2018

International Organization for Migration (2018) Data bulletin-big data and migration. https://publications.iom.int/system/files/pdf/issue_5_big_data_and_migration.pdf. Accessed 23 June 2018

Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28(5):26

Lu X, Bengtsson L, Holme P (2012) Predictability of population displacement after the 2010 Haiti earthquake. Proc Natl Acad Sci USA 109(29):11576–81

Lu X, Wrathall DJ, Sundsoy PR, Nadiruzzaman M, Wetter E, Iqbal A, Qureshi T, Tatem A, Canright G, Engo-Monsen K, Bengtsson L (2016) Unveiling hidden migration and mobility patterns in climate stressed regions: a longitudinal study of six million anonymous mobile phone users in Bangladesh. Glob Environ Change-Human Policy Dimens 38:1–7

Mobile Telecommunications (2018) Coverage. http://www.mtc.com.na/coverage. Accessed 3 June 2018

Namibia Statistics Agency (2013) Namibia 2011 Population and Housing Census Main Report. https://cms.my.na/assets/documents/p19dmn58guram30ttun89rdrp1.pdf. Accessed 22 Nov 2017

Namibia Statistics Agency (2015) Namibia 2011 Census Migration Report. https://cms.my.na/assets/documents/Migration_Report.pdf. Accessed 22 Nov 2017

Pendleton W, Crush J, Nickanor N (2014) Migrant windhoek: rural–urban migration and food security in Namibia. Urban Forum 25(2):191–205

R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/

Ruktanonchai NW, Bhavnani D, Sorichetta A, Bengtsson L, Carter KH, Cordoba RC, Le Menach A, Lu X, Wetter E, Erbach-Schoenberg EZ, Tatem AJ (2016a) Census-derived migration data as a tool for informing malaria elimination policy. Malaria J 15:273

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

Ruktanonchai NW, Ruktanonchai CW, Floyd JR, Tatem AJ (2018) Using Google Location History data to quantify fine-scale human mobility. Int J Health Geogr 17:28

Ruktanonchai A, Bird TJ, Ruktanonchai NW, Erbach-Schoenberg EZ, Pezzulo C, Tejedor N, Waldock IC, Sadler JD, Garcia AJ, Sedda L, Tatem AJ (2016) Mapping internal connectivity through human migration in malaria endemic countries. Scientific Data 3:160066

Stork C (2011) Namibian Sector Performance 2011. Research ICT Africa. http://www.researchictafrica.net/publications/Evidence_for_ICT_Policy_Action/Stork_C_-_2011_Namibian_Sector_Performance_Review.pdf. Accessed 13 Nov 2017

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

The GSM Association (2018) The Mobile Economy 2018. https://www.gsma.com/mobileeconomy/. Accessed 30 July 2018

The Namibia Ministry of Health, Social Services-MoHSS/Namibia and ICF International (2014) Namibia Demographic and Health Survey 2013. MoHSS/Namibia and ICF International, Windhoek, Namibia

Tobler WR (1970) A computer movie simulating urban growth in the detroit region. Econ Geogr 46(2):234–240

Vobruba T, Korner A, Breitenecker F (2016) Modelling, analysis and simulation of a spatial interaction model. Ifac Pap 49(29):221–225

Wardrop NA, Jochem WC, Bird TJ, Chamberlain HR, Clarke D, Kerr D, Bengtsson L, Juran S, Seaman V, Tatem AJ (2018) Spatially disaggregated population estimates in the absence of national population and housing census data. Proc Natl Acad Sci USA 115(14): 3529–3537

Wesolowski A, Buckee CO, Bengtsson L, Wetter E, Lu X, Tatem AJ (2014a) Commentary: containing the ebola outbreak–the potential and challenge of mobile network data. PLoS Curr 6, http://currents.plos.org/outbreaks/index.html%3Fp=42561.html

Wesolowski A, Buckee CO, Pindolia DK, Eagle N, Smith DL, Garcia AJ, Tatem AJ (2013) The use of census migration data to approximate human movement patterns across temporal scales PloS ONE 8(1):e52971

Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, Buckee CO (2012) Quantifying the impact of human mobility on malaria. Science 338 (6104):267–70

Wesolowski A, Metcalf CJE, Eagle N, Kombich J, Grenfell BT, Bjornstad ON, Lessler J, Tatem AJ, Buckee CO (2015a) Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. Proc Natl Acad Sci USA 112(35):11114–11119

Wesolowski A, O'Meara WP, Eagle N, Tatem AJ, Buckee CO (2015b) Evaluating spatial interaction models for regional mobility in sub-Saharan Africa. PLoS Comput Biol 11(7):e1004267

Wesolowski A, Qureshi T, Boni MF, Sundsoy PR, Johansson MA, Rasheed SB, Engo-Monsen K, Buckee CO (2015c) Impact of human mobility on the emergence of dengue epidemics in Pakistan. Proc Natl Acad Sci USA 112 (38):11887–11892

Wesolowski A, Stresman G, Eagle N, Stevenson J, Owaga C, Marube E, Bousema T, Drakeley C, Cox J, Buckee CO (2014b) Quantifying travel behavior for infectious disease research: a comparison of data from surveys and mobile phones. Sci Rep 4:5678

Wesolowski A, Zu Erbach-Schoenberg E, Tatem AJ, Lourenco C, Viboud C, Charu V, Eagle N, Engo-Monsen K, Qureshi T, Buckee CO, Metcalf CJE (2017) Multinational patterns of seasonal asymmetry in human movement influence infectious disease dynamics. Nat Commun 8(1):2069

Zhao D, Li ZJ, Zhou H, Lai SJ, Yin WW, Yang WZ (2012) [Review on the research progress of early-warning system on dengue fever]. Zhonghua Liu Xing Bing Xue Za Zhi 33(5):540–3

Zipf GK (1946) The P1 P2/D hypothesis: on the intercity movement of persons. Am Sociol Rev 11(6):677–686

Zu Erbach-Schoenberg E, Alegana VA, Sorichetta A, Linard C, Lourenco 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 Metr 14:35

Author contributions

S.L., E.z.E.-S., and A.J.T. conceived and designed the manuscript; S.L., E.z.E.-S., and C.P. performed the analysis; S.L., E.z.E.-S., C.P., N.W.R., A.S., J.S., T.L., C.A.D., and A.J.T. wrote the paper.

Additional information

The online version of this article (https://doi.org/10.1057/s41599-019-0242-9) contains supplementary material, which is available to authorized users.

Competing interests: The authors declare no competing interests.