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*]

[1]WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton SO17 1BJ, United Kingdom

[2]Flowminder Foundation, SE-113 55 Stockholm, Sweden

[3]School of Public Health, Fudan University, Key Laboratory of Public Health Safety, Ministry of Education, 130 Dongan Road, Shanghai 200032, China

Shengjie Lai and Elisabeth zu Erbach-Schoenberg contributed equally to this work.

*Correspondence and requests for materials should be addressed to A.J.T (email: Andy.tatem@gmail.com) or S.L. (email: laishengjie@foxmail.com).

ABSTRACT

Statistics on internal migration are important for keeping estimates of subnational 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. Analysing 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 modelled 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 to censuses. Moreover, we show how gravity-type spatial interaction models built using CDRs can accurately capture migration flows. 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.

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 one- or five-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 modelling of migration flows emerged, leading to the advanced development of modelling 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 analysing 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 temporarily explicit estimates of human mobility 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. Additionally, 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, labour 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:

$$MIG_{i,j} = \beta_0 + \beta_1 CDR_{i,j} + \vec{\beta}[X] \quad [1]$$

where the dependent variable $MIG_{i,j}$ is comprised of the observed migration flows between regions in Namibia from the census. $CDR_{i,j}$ 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 $\vec{\beta}$.

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:

$$MIG_{i,j} = \frac{POP_i^{\beta_1} POP_j^{\beta_2}}{DIST_{i,j}^{\beta_3}} \quad [2]$$

where POP_i and POP_j refer to populations at an origin i and a destination j in 2010, respectively; $DIST_{i,j}$ 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:

$$\frac{MIG_{i,j}}{TOT_i} = \frac{e^{\beta_0+\beta_1 P_i+\beta_2 P_j-\beta_3 DIST_{i,j}}}{1+e^{\beta_0+\beta_1 P_i+\beta_2 P_j-\beta_3 DIST_{i,j}}} \quad [3]$$

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- 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 Figs. 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. 2*A*). 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- 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 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 to GTSIMs and CGTSIMs (Fig. 3). Additionally, in terms of outflow, inflow, and net 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 (Figs. 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 respectively to adjust CDRs. The performance of both CDRLMs and CGTSIMs were not significantly improved by these adjustments however (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 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 analysing 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 up to 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 to 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- 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 modelling 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 behaviour 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.

Additionally, 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. Additionally, 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 analysing 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.

Figures

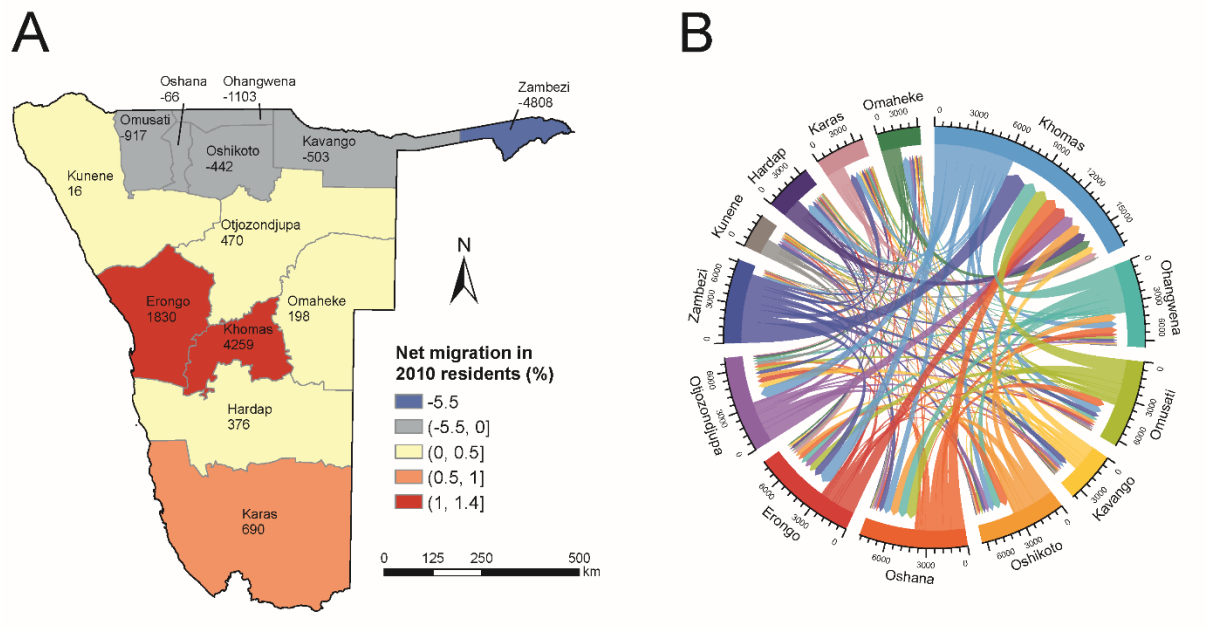

Fig. 1. Census-derived internal migration in Namibia, September 2010 – August 2011. (*A*) Net migration by region. The number of net migrants by region is presented under the name of each region. (*B*) Circular plot of migrant flows between regions. The origins and destinations of migrants are each assigned a colour and represented by the circle's segments. The direction of the flow is encoded by both the origin region's colour 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).

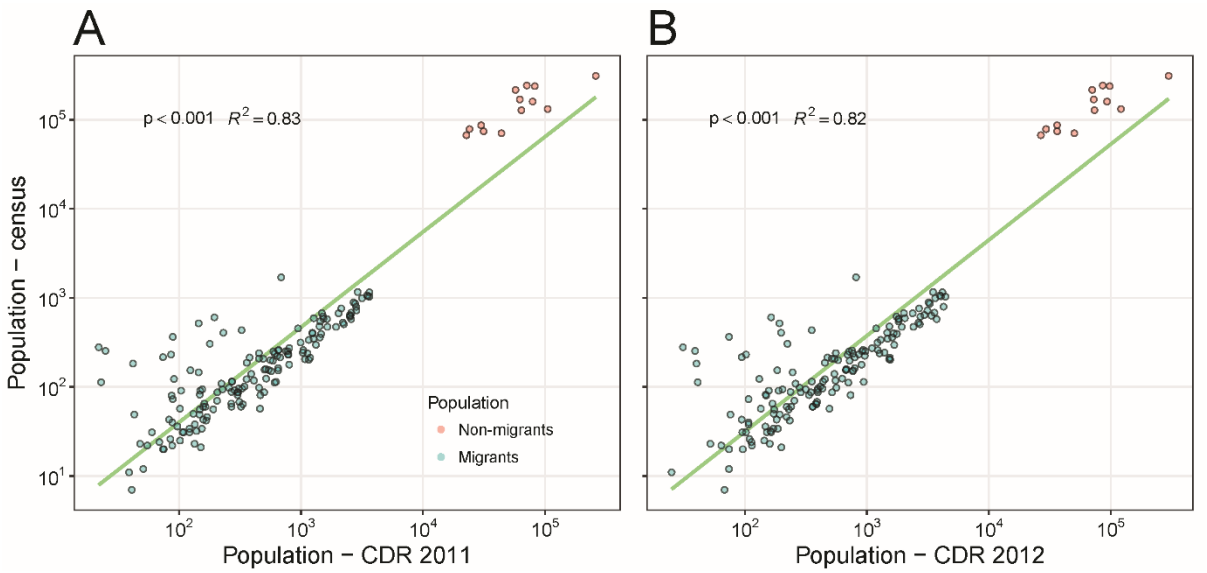

Fig. 2. Logarithmic relation between census-derived populations in 2011 and the number of CDR-derived mobile phone users in Periods 2011 (*A*) and 2012 (*B*) at regional level. The green solid lines represent linear regression fit, with *p* and R^2 values provided.

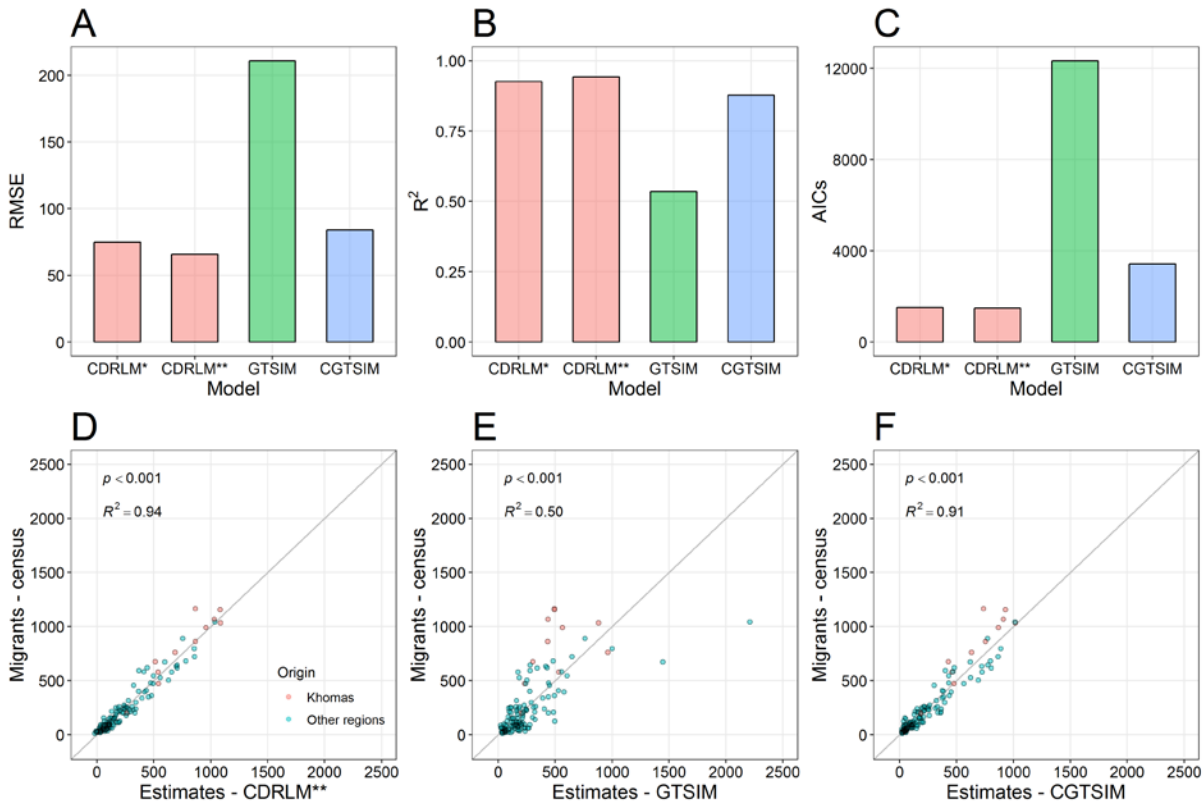

Fig. 3. Precision and accuracy assessments of models for replicating 2011 census-derived migration statistics. The indicators of (*A*) root-mean-square error (RMSE), (*B*) R^2, and (*C*) Akaike Information Criterion (AIC), were computed to compare three types of models: CDR-based linear model (CDRLM), 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.

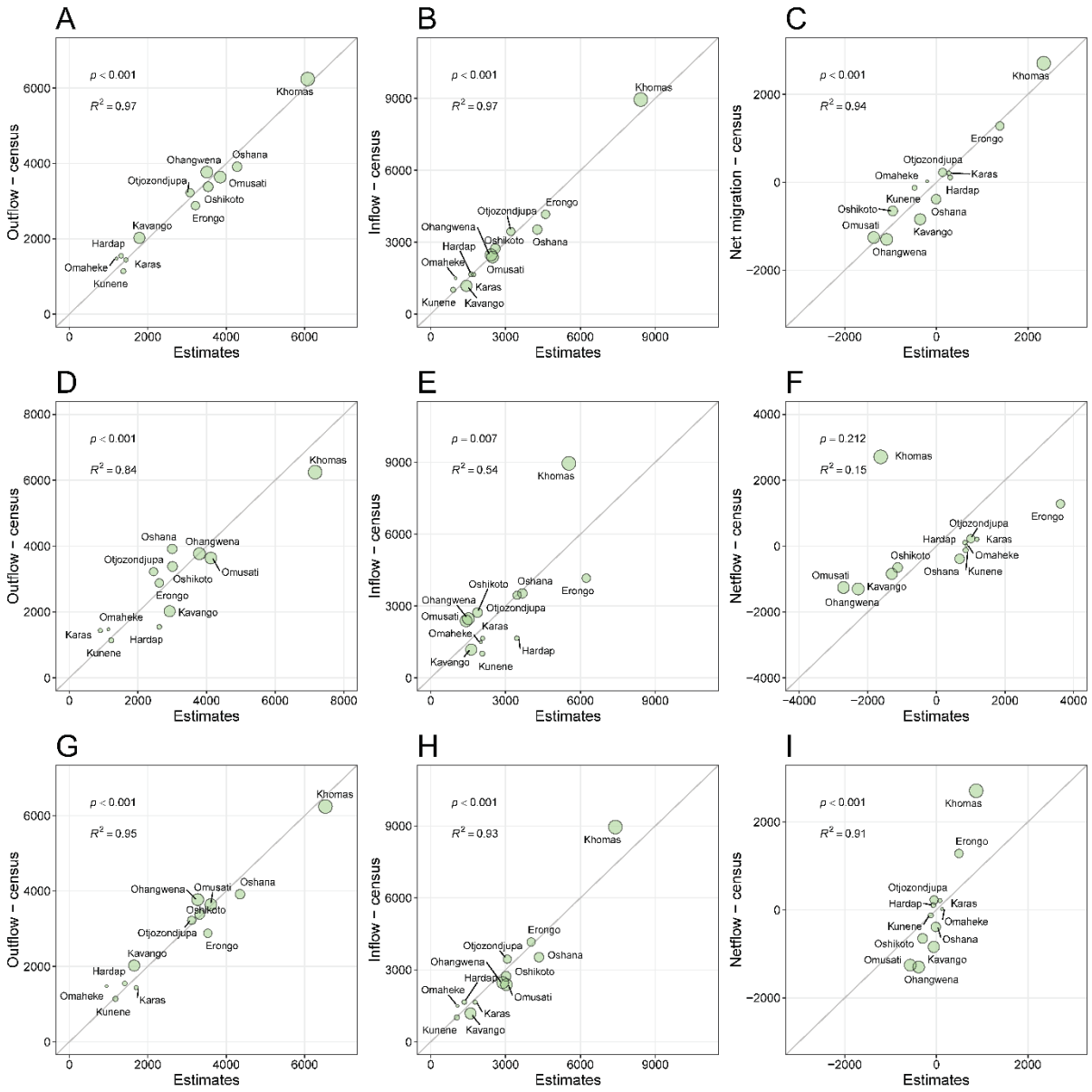

Fig. 4. Comparing regional outflow, inflow and net migration between census data and models' estimates for Period 2011: CDRLM in (*A*), (*B*) and (*C*), GTSIM in (*D*), (*E*) and (*F*), and CGTSIM in (*G*), (*H*) and (*I*). For each type of model, the results of model with the lowest RMSE are presented. The Zambezi region as an outliner is excluded, and unadjusted CDR data are used.

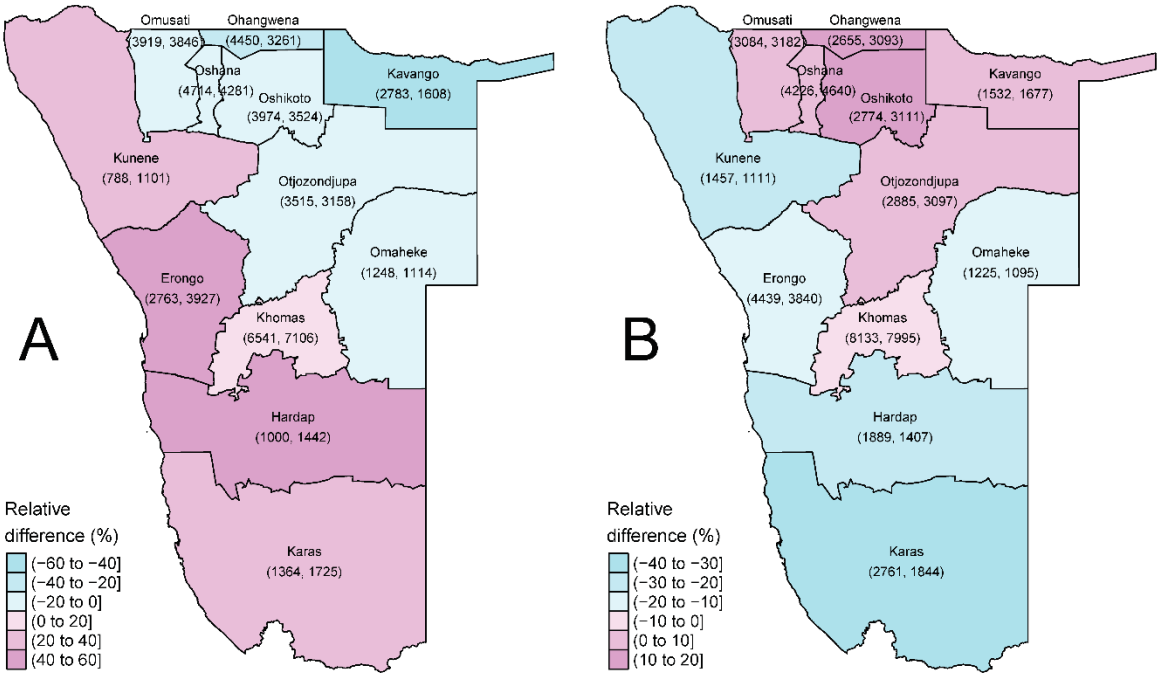

Fig. 5. Relative difference of regional outflow (*A*) and inflow (*B*) between Period 2011 and Period 2012. The migrations were estimated by the CDRLM using only CDRs, and the adjusted CDR data of Period 2012 were used to offset the impact of the increasing mobile phone ownership across periods. The 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.

Conflicts of interest

The authors declare that they have no conflicts.

Acknowledgements

The authors would like to thank MTC for providing access to the mobile phone data. S.L. is supported by the grants from the National Natural Science Fund (No. 81773498), the Ministry of Science and Technology of China (2016ZX10004222-009), and the Program of Shanghai Academic/Technology Research Leader (No. 18XD1400300). A.J.T. is supported by funding from the Bill & Melinda Gates Foundation (OPP1106427, 1032350, OPP1134076, OPP1094793), the Clinton Health Access Initiative, the UK Department for International Development (DFID) and the Wellcome Trust (106866/Z/15/Z, 204613/Z/16/Z). C.P. is supported by funding from the Bill & Melinda Gates Foundation (OPP1134076). J.S. is supported by funding from the Belgian Federal Science Policy Office (BELSPO). This work forms part of the outputs of WorldPop (www.worldpop.org) and Flowminder (www.flowminder.org). The funders had no role in the study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

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 & 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 analysed 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.

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.

Reference

Abel G J (2013) Estimating global migration flow tables using place of birth data. *Demographic Research*; **28**: 505-546.

Abel G J and Sander N (2014) Quantifying global international migration flows. *Science*; **343**(6178): 1520-2.

Agliari E, Barra A, Contucci P, Pizzoferrato A and Vernia C (2018) Social interaction effects on immigrant integration. *Palgrave Communications*; **4**

Blumenstock J E (2012) Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda. *Information Technology for Development*; **18**(2): 107-125.

Cohen J E, Roig M, Reuman D C and GoGwilt C (2008) International migration beyond gravity: a statistical model for use in population projections. *Proc Natl Acad Sci U S A*; **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 F R, Gaughan A E, Blondel V D and Tatem A J (2014) Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*; **111**(45): 15888-15893.

Frayne B (2005) Survival of the poorest: migration and food security in Namibia. *Agropolis. The social, political and environmental dimensions of urban agriculture*: 31-44.

Garcia A J, Pindolia D K, Lopiano K K and Tatem A J (2015) Modeling internal migration flows in sub-Saharan Africa using census microdata. *Migration Studies*; **3**(1): 89-110.

Getis A (2008) A history of the concept of spatial autocorrelation: A geographer's perspective. *Geographical Analysis*; **40**(3): 297-309.

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

Henry S, Boyle P and Lambin E F (2003) Modelling inter-provincial migration in Burkina Faso, West Africa: the role of socio-demographic and environmental factors. *Applied Geography*; **23**(2-3): 115-136.

Henry S, Piche V, Ouedraogo D and Lambin E F (2004) Descriptive analysis of the individual migratory pathways according to environmental typologies. *Population and Environment*; **25**(5): 397-422.

Hua C-i and Porell F (1979) A Critical Review of the Development of the Gravity Model. *International Regional Science Review*; **4**(2): 97-126.

International Federation of Red Cross and Red Crescent Societies (2011) *DREF operation final report: Namibia Floods*. Available from: 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. Available from: 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*. Available from: 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*. Available from: 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*. Available from: 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*. Available from: https://publications.iom.int/system/files/pdf/issue_5_big_data_and_migration.pdf [Accessed 23 June]

Kuhn M (2008) Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*; **28**(5): 26.

Lu X, Bengtsson L and Holme P (2012) Predictability of population displacement after the 2010 Haiti earthquake. *Proc Natl Acad Sci U S A*; **109**(29): 11576-81.

Lu X, Wrathall D J, Sundsoy P R, Nadiruzzaman M, Wetter E, Iqbal A, Qureshi T, Tatem A, Canright G, Engo-Monsen K and 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. *Global Environmental Change-Human and Policy Dimensions*; **38**: 1-7.

Mobile Telecommunications (2018) *Coverage*. Available from: http://www.mtc.com.na/coverage [Accessed 3 June 2018]

Namibia Statistics Agency (2013) *Namibia 2011 Population & Housing Census Main Report*. Available from: https://cms.my.na/assets/documents/p19dmn58guram30ttun89rdrp1.pdf [Accessed 22 November 2017]

Namibia Statistics Agency (2015) *Namibia 2011 Census Migration Report*. Available from: https://cms.my.na/assets/documents/Migration_Report.pdf [Accessed 22 November 2017]

Pendleton W, Crush J and 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.* Available from: https://www.R-project.org/]

Ruktanonchai N W, Bhavnani D, Sorichetta A, Bengtsson L, Carter K H, Cordoba R C, Le Menach A, Lu X, Wetter E, Erbach-Schoenberg E Z and Tatem A J (2016a) Census-derived migration data as a tool for informing malaria

elimination policy. *Malaria Journal*; **15**

Ruktanonchai N W, DeLeenheer P, Tatem A J, Alegana V A, Caughlin T T, Erbach-Schoenberg E Z, Lourenco C, Ruktanonchai C W and Smith D L (2016b) Identifying Malaria Transmission Foci for Elimination Using Human Mobility Data. *Plos Computational Biology*; **12**(4)

Ruktanonchai N W, Ruktanonchai C W, Floyd J R and Tatem A J (2018) Using Google Location History data to quantify fine-scale human mobility. *International Journal of Health Geographics*; **17**

Sorichetta A, Bird T J, Ruktanonchai N W, Erbach-Schoenberg E Z, Pezzulo C, Tejedor N, Waldock I C, Sadler J D, Garcia A J, Sedda L and Tatem A J (2016) Mapping internal connectivity through human migration in malaria endemic countries. *Scientific Data*; **3**

Stork C (2011) *Namibian Sector Performance 2011*. Research ICT Africa. Available from: http://www.researchictafrica.net/publications/Evidence_for_ICT_Policy_Action/Stork_C_-_2011_Namibian_Sector_Performance_Review.pdf [Accessed 13 November 2017]

Tatem A J, Huang Z, Narib C, Kumar U, Kandula D, Pindolia D K, Smith D L, Cohen J M, Graupe B, Uusiku P and 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*. Available from: https://www.gsma.com/mobileeconomy/ [Accessed 30 July]

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

Tobler W R (1970) A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*; **46**(2): 234-240.

Vobruba T, Korner A and Breitenecker F (2016) Modelling, Analysis and Simulation of a Spatial Interaction Model. *Ifac Papersonline*; **49**(29): 221-225.

Wardrop N A, Jochem W C, Bird T J, Chamberlain H R, Clarke D, Kerr D, Bengtsson L, Juran S, Seaman V and Tatem A J (2018) Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences of the United States of America*; **115**(14): 3529-3537.

Wesolowski A, Buckee C O, Bengtsson L, Wetter E, Lu X and Tatem A J (2014a) Commentary: containing the ebola outbreak - the potential and challenge of mobile network data. *PLoS Curr*; **6**

Wesolowski A, Buckee C O, Pindolia D K, Eagle N, Smith D L, Garcia A J and Tatem A J (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 A J, Smith D L, Noor A M, Snow R W and Buckee C O (2012) Quantifying the impact of human mobility on malaria. *Science*; **338**(6104): 267-70.

Wesolowski A, Metcalf C J E, Eagle N, Kombich J, Grenfell B T, Bjornstad O N, Lessler J, Tatem A J and Buckee C O (2015a) Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*; **112**(35): 11114-11119.

Wesolowski A, O'Meara W P, Eagle N, Tatem A J and Buckee C O (2015b) Evaluating Spatial Interaction Models for Regional Mobility in Sub-Saharan Africa. *Plos Computational Biology*; **11**(7)

Wesolowski A, Qureshi T, Boni M F, Sundsoy P R, Johansson M A, Rasheed S B, Engo-Monsen K and Buckee C O (2015c) Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proceedings of the National Academy of Sciences of the United States of America*; **112**(38): 11887-11892.

Wesolowski A, Stresman G, Eagle N, Stevenson J, Owaga C, Marube E, Bousema T, Drakeley C, Cox J and Buckee C O (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 A J, Lourenco C, Viboud C, Charu V, Eagle N, Engo-Monsen K, Qureshi T, Buckee C O and Metcalf C J E (2017) Multinational patterns of seasonal asymmetry in human movement influence infectious disease dynamics. *Nature Communications*; **8**(1): 2069.

Zhao D, Li Z J, Zhou H, Lai S J, Yin W W and Yang W Z (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 G K (1946) The P1 P2/D Hypothesis: On the Intercity Movement of Persons. *American Sociological Review*; **11**(6): 677-686.

Zu Erbach-Schoenberg E, Alegana V A, Sorichetta A, Linard C, Lourenco C, Ruktanonchai N W, Graupe B, Bird T J, Pezzulo C, Wesolowski A and Tatem A J (2016) Dynamic denominators: the impact of seasonally varying population numbers on disease incidence estimates. *Popul Health Metr*; **14**: 35.

Supporting Information for

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*]

[1]WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton SO17 1BJ, United Kingdom

[2]Flowminder Foundation, SE-113 55 Stockholm, Sweden

[3]School of Public Health, Fudan University, Key Laboratory of Public Health Safety, Ministry of Education, 130 Dongan Road, Shanghai 200032, China

Shengjie Lai and Elisabeth zu Erbach-Schoenberg contributed equally to this work.

*Correspondence and requests for materials should be addressed to A.J.T (email: Andy.tatem@gmail.com) or S.L. (email: laishengjie@foxmail.com).

This PDF file includes:

SI Text

1.1. CDR Data Processing

Each call detail record (CDR) contained the location of the tower that the communication was routed through, giving us an approximate estimate of the user's location. Since we aimed for region level estimates, we mapped each entry to the region where the routing tower was contained (Mobile Telecommunications 2018). The daily location of a mobile phone user could be defined if the user made or received at least one communication (call or text) on that day. To reduce occasional travel or seasonal movement, e.g. travel over the Christmas holidays, infrequent mobile phone users with 30 days or less worth of defined daily locations (next section) for each year (12 months) were filtered out. This corresponded to 12% mobile phone users that were filtered out between October 2010 and September 2012 (Period 2011), and 20% users between October 2011 and September 2013 (Period 2012), respectively. The higher proportion of users filtered out in Period 2012 may be due to the constant increase of new SIM subscribers and shifting mobile phone usage from voice to data (Mobile Telecommunications 2012, 2014). We then calculated the most frequently observed region for each user and day as that user's daily location. Using these daily locations, we could define the residence as the most frequent daily location during a given year-long period (12 months) for each user. Finally, we derived migration flows of mobile users by comparing the places of residence at regional level across years for each user.

1.2. Comparing Different Time Lengths to Define Residence

As sufficient CDR data points are needed to accurately estimate the place of residence, different time windows of data were compared to define the residences of mobile phone users. We first investigated how many months of data were available for each mobile phone user in an example dataset from January to December 2012 that was used for migration estimate in Periods 2011 and 2012. A monthly location for each individual was calculated as the most frequent daily location on regional levels over each month if this user made at least one call or text in that month. We found that the most frequent case was that users have data available for all months of the year (Figure S2A), but it still showed a high proportion (15%) of infrequent users with only 1-month of data, which could introduce a strong bias for deriving migration flows

compared to using yearly locations as residence. Therefore, to exclude very infrequent users, we used 31 days of defined daily locations as a cutoff for an individual to be included in the migration estimation. Moreover, most individuals with more than 1-month of data had monthly locations that were identical to yearly locations (Figure S2B), and the spatial differences in the mean percentage of monthly locations matching yearly locations were likely homogenous across the country, from the lowest percentage of 91% to the highest of 97% (Figure S2C).

Seasonal movements might lead to an individual temporarily residing at different locations, which would influence estimates of the place of residence in settings where only shorter periods (e.g. a month or two per year) of CDRs are available (Wesolowski et al., 2017). Figure S3A shows the percentage of locations using the random N months (ranged 1-11 months) of data that match the locations using a full year of data. As expected, the accuracy of estimating the usual residence increases with increasing length of period covered by the data. However, we observed the largest increase going from one month to two months of data used, likely due to the strong seasonal effect of individuals' temporary locations deviating from their usual places of residence during holiday periods and other times when short-term mobility is prevalent. Furthermore, Figure S3B shows the Z-score of the percentage of users whose monthly locations matched the yearly location. The negative deviation in December and January means fewer users being found at their usual residences in these holiday months, with a similar, but smaller effect being seen in May. This is part of the reason that censuses are generally timed to occur during a period with low seasonal population movement, e.g. August and September, and again it highlights the need to exclude infrequent mobile phone users to prevent the inclusion of short-term travel in longer-term migration estimation.

1.3. Phone Ownership

As mobile phone ownership is not homogenous across the population, we utilized data from the 2013 Namibia Demographic and Health Survey (DHS) (The Namibia Ministry of Health et al., 2014) to assess to what extent there is a possible exclusion of certain groups from the CDRs, and the potential of using the indicators to adjust CDR-based migration estimates and thus, to better match the census-based observations. Using a stratified two-stage cluster design, the 2013 Namibia DHS sampled 9,849 households that were distributed in 554 clusters across the country, with 269 in urban

areas and 285 in rural areas (The Namibia Ministry of Health et al., 2014). This survey collected data for a wide range of indicators including demographic, health and socio-economic characteristics, and it also reported information on mobile phone ownership at household level. Moreover, households were ranked by the International Wealth Indicator (IWI) score, indicating to what extent the household possesses a basic set of assets. The IWI score is an asset-based wealth index used for measuring the economic situation of households in developing countries (Smits and Steendijk 2015). Households were then divided by five quintiles, based on the IWI ranking. Households falling within the upper two quintiles ('richer' and 'richest') were then classified as being wealthy, while households in the bottom three quintiles were defined as not being wealthy.

We first performed an exploratory bivariate analysis to compare the characteristics between households owning at least one mobile phone and households without mobile phones. Data were weighted using sampling weights and adjusted for the survey sampling design, and analyses were done using STATA 14 software. Table S2 shows that households owning at least one mobile phone are significantly different from households without mobile phones for almost all background characteristics considered in this analysis. In comparisons to households without mobile phones, those with mobiles are significantly more likely to be wealthy, to reside in urban areas and live in Khomas (22%). Households without mobile phones are more likely to be located in Kavango, Zambezi and Ohangwen regions. Moreover, households with one or more mobile phones are significantly more likely to have younger heads and a higher number of household members, whereas households without mobile phones tend to have more uneducated household heads and uneducated female and male residents (Table S2).

Furthermore, we performed a binary logistic regression model for the probability of households without mobile phones, by including all variables that resulted as being statistically significant in the bivariate analysis. In particular, we aimed to identify groups of households that had a high probability of not owning a mobile phone and the characteristics they share within a multiple regression analysis framework, after adjusting for the effects of other variables (Callegaro and Poggio 2004). Table S3 shows that there is a significant differential in the ownership of mobiles between households regarding wealth, age of the household head, household size, and education. Our findings also show that there is a significant ownership differential

between regions in Namibia, confirming the results from the bivariate analysis. The odds of ownership of a mobile phone for households residing in most regions range between about 2 and 5 times greater than that in Kavango, meaning it may be necessary to take the regional mobile ownership bias into account in estimates of migration by CDRs for each region.

1.4. Model Covariates

We also collated potential migration-related demographic, socioeconomic, geographic and environmental variables for migration modelling as described in previous studies (Henry et al., 2003; Henry et al., 2004; Garcia et al., 2015; Wesolowski et al., 2015; Ruktanonchai et al., 2016; Sorichetta et al., 2016; Vobruba et al., 2016). An administrative unit boundary file at regional level matching the year of the census was obtained from the Global Administrative Areas Database (GADM 2018). Following previous studies (Garcia et al., 2015; Sorichetta et al., 2016), the shapefile was used to calculate variables that measure distance and contiguity between administrative units, respectively. Euclidean distance between geometric centroids is commonly used as a parameter in gravity models, where it represents the barriers to, as well as potential costs of, migration (Garcia et al., 2015; Sorichetta et al., 2016). To calculate possible environmental drivers of migration in models, moreover, high resolution monthly precipitation grids (30 seconds, ~1 km^2) were obtained from WorldClim version 2 (worldclim.org/version2). We then aggregated the precipitation data to obtain average annual precipitation (mm) by region as a proxy of push-pull factors such as agricultural productivity and the potential of floods and droughts.

A variety of demographic and socioeconomic variables known to be associated with migration flows were also collated from 2011 census data for each region in Namibia (Table S1). First, we included the populations in origin and destination (POP_i and POP_j) in 2010 and 2011. Given that the urbanization can be a significant pull factor for migrants (Lall et al., 2006; Garcia et al., 2015), then we included the percentage of population living in urban areas in origin and destination, denoted as $URBAN_i$ and $URBAN_j$, respectively.

Previous studies on migration also suggested that human migration is, at least in part, driven by economic opportunities and that different demographic characteristics, such as age, sex, educational attainment and marital status influence

migration rates (Henry et al., 2004; Garcia et al., 2015; Sorichetta et al., 2016). Therefore, we also collated the following covariates from 2011 Namibia census data: *MALEPROP*, the proportion of males in the region; *AGE15_19*, the proportion of the population between 15 and 59 years old in the region; *ACTIVE*, the proportion of labour force participation in the population aged 15 and above by region; *LITERACY*, the percentage of literate individuals out of the population aged 15 and above; and *SINGLEPROP*, the proportion of unmarried people in the population aged 15 and above (Namibia Statistics Agency 2013). However, all of these variables show strong correlations (Pearson's r >0.5) with $URBAN_i$ and $URBAN_j$, in part because urban areas might offer more opportunities for improving these socioeconomic status (Lall et al., 2006). As the urbanization is related to socioeconomic development, we removed the demographic and socioeconomic variables that were highly correlated with the urbanization variables to avoid multicollinearity and overfitting in models.

SI Tables

Table S1. The summary of models.

Type	Model	Independent variables	
		Mobile phone data	**Other variables**
CDR-based linear model (CDRLM)	1	$CDR_{i,j}$ or $adjCDR_{i,j}$ or $CDR_{i,j} + PHONEPROP_i$	None
	2		$+ POP_i + POP_j + DIST_{i,j}$
	3[a]		$+ URBAN_i + URBAN_j + DIST_{i,j}$
	4[b]		$+ RAIN_i + RAIN_j + DIST_{i,j}$
Gravity-type spatial interaction model (GTSIM)	1	None	$\ln(POP_i) + \ln(POP_j) + DIST_{i,j}$
	2[a,b]		$URBAN_i + URBAN_j + DIST_{i,j}$
	3		$RAIN_i + RAIN_j + DIST_{i,j}$
CDR-based Gravity-type spatial interaction model (CGTSIM)	1[a,b]	$CDR_{i,j}$ or $adjCDR_{i,j}$ or $CDR_{i,j} + PHONEPROP_i$	$+ \ln(POP_i) + \ln(POP_j) + DIST_{i,j}$
	2		$+ URBAN_i + URBAN_j + DIST_{i,j}$
	3		$+ RAIN_i + RAIN_j + DIST_{i,j}$

Variable descriptions:

$CDR_{i,j}$: Number of migrating mobile phone users from origin i to destination j based on call detail records (CDRs).

$PHONEPROP_i$: Proportion of the population owning mobile phones in origin region i.

$adjCDR_{i,j}$: $CDR_{i,j}$ divided by $PHONEPROP_i$.

POP_i and POP_j : Population of origin i and destination j.

$URBAN_i$ and $URBAN_j$: Proportion of population living in urban areas.

$RAIN_i$ and $RAIN_j$: Annual average precipitation (mm).

$DIST_{i,j}$: Euclidean distance between centroids of origin i and destination j.

[a] Optimal model of each model family for regions except Zambezi, based on the lowest root-mean-square error (RMSE) (Fig. S8).

[b] Optimal model of each model family for all regions, based on the lowest RMSE (Fig. S8).

Table S2. Demographic and socio-economic characteristics of households by mobile phone ownership.

	With MP % (N=8,589)	Without MP % (N=1,257)	Total % (N=9,846)	p value*
Gender of household head				0.159
Male	55.7	58.6	56.1	
Female	44.3	41.4	43.9	
Mean age of household head, yrs (SE)	45.5 (0.32)	50.7 (0.76)		<0.001**
Mean number of household members, yrs (SE)	4.4 (0.05)	3.6 (0.12)		<0.001**
Number of under-five children in households				0.019
0	56.2	62.7	57.0	
1-3	41.9	36.1	41.3	
≥4	1.8	1.2	1.8	
Ratio of under 5 yrs in household members (SE)	0.120 (0.002)	0.113 (0.006)		0.257
Household wealth (International Wealth Index, IWI)				<0.001
Non-wealthy	52.4	91.4	56.8	
Wealthy	47.5	8.5	43.1	
Household residence				<0.001
Urban	55.8	22.3	52	
Rural	44.2	77.7	48	
Region				<0.001
Zambezi (Caprivi)	4.8	10.6	5.5	
Erongo	10.1	4.2	9.5	
Hardap	3.8	4.4	3.9	
Karas	4.3	3	4.1	
Kavango	6.2	17.6	7.5	
Khomas	22.1	7.5	20.4	
Kunene	3.1	7.3	3.6	
Ohangwena	8.8	11.5	9.1	
Omaheke	3.2	4.9	3.4	
Omusati	9.7	9.4	9.6	
Oshana	8.8	6	8.4	
Oshikoto	8.3	8.3	8.3	
Otjozondjupa	6.8	5.4	6.6	
Education attainment of household head* **				<0.001
No education	13	36.3	15.7	
Incomplete primary	21.4	29.8	22.4	
Complete primary	5.5	8.3	5.8	
Incomplete secondary	30.6	18.1	29.2	
Complete secondary	16.3	4.9	15	
Higher	12.6	2.2	11.4	
Unknown	0.6	0.4	0.6	
Education attainment of female household population (aged 6+)^				<0.001

	With MP % (N=8,589)	Without MP % (N=1,257)	Total % (N=9,846)	p value*
No education	10.6	28.9	12.2	
Incomplete primary	32	42.2	32.9	
Complete primary	5.5	6.3	5.6	
Incomplete secondary	32.4	18	31.1	
Complete secondary	11.9	3.1	11.1	
Higher	7.4	1.4	6.8	
Unknown	0.2	0.1	0.2	
Education attainment of male household population (aged 6+)^^				<0.001
No education	12	28	13.6	
Incomplete primary	35.7	41.2	36.2	
Complete primary	5.3	6.8	5.4	
Incomplete secondary	28	19	27.1	
Complete secondary	11.2	3.6	10.4	
Higher	7.5	1.2	6.8	
Unknown	0.3	0.3	0.3	

MP: mobile phone. The data were obtained from the Namibia 2013 Demographic and Health Survey (DHS) (dhsprogram.com/pubs/pdf/fr298/fr298.pdf). The ownership was defined as the presence of at least one mobile phone in the household. Percentages were produced using sampling weights and adjusted for the survey design. *Pearson*'s Chi-square test and t-test (where specified) were run to test the association between variables. Where indicated, variables were constructed using the household member dataset.

* *p* values were produced using Pearson's Chi-square test, adjusted for the survey design.

** Calculated using a t-test for equality of means.

*** Information available for total population, N=9,796.

^ calculated for de facto female household members aged 6 and over by highest level of schooling attended or completed (N=18,230).

^^ calculated for de facto male household members aged 6 and over by highest level of schooling attended or completed (N=16,153).

Table S3. Estimates of factors for household without mobile phone (N=9,842).

	Coefficient	SE	Odds Ratio (95% CI)
Region (Ref.: Kavango)			
Zambezi (Caprivi)	-0.46	0.219	0.63 (0.41; 0.97)*
Erongo	-1.12	0.267	0.33 (0.19; 0.55)*
Hardap	-0.44	0.224	0.64 (0.41; 1.00)
Karas	-1.06	0.333	0.35 (0.18; 0.67)*
Khomas	-0.92	0.232	0.40 (0.25; 0.63)*
Kunene	-0.38	0.226	0.69 (0.44; 1.07)
Ohangwena	-1.09	0.264	0.34 (0.20; 0.56)*
Omaheke	-0.83	0.236	0.44 (0.28; 0.70)*
Omusati	-1.61	0.232	0.20 (0.13; 0.31)*
Oshana	-1.35	0.424	0.26 (0.11; 0.60)*
Oshikoto	-1.30	0.223	0.27 (0.18; 0.42)*
Otjozondjupa	-0.99	0.276	0.37 (0.22; 0.64)*
Place of residence (Ref.: Urban)			
Rural	0.80	0.137	2.24 (1.71; 2.93)*
Household wealth quintiles IWI (Ref.: Wealthy)			
Not wealthy	1.50	0.138	4.47 (3.41; 5.86)*
Household has at least one member with completed secondary or higher education (Ref.: Yes)			
No	0.84	0.12	2.32 (1.84; 2.94)*
Age of household head	0.02	0.003	1.02 (1.01; 1.02)*
Number of household members	-0.19	0.021	0.83 (0.79; 0.86)*
Intercept	-3.50	0.238	0.03 (0.02; 0.05)*

Note: A logistic regression model was used, and the estimates of coefficients, odds ratios and 95% confidence intervals (CI) are presented. “Svy” command in STATA was used to account for weighting and sampling design. * p value < 0.05.

SI Figures

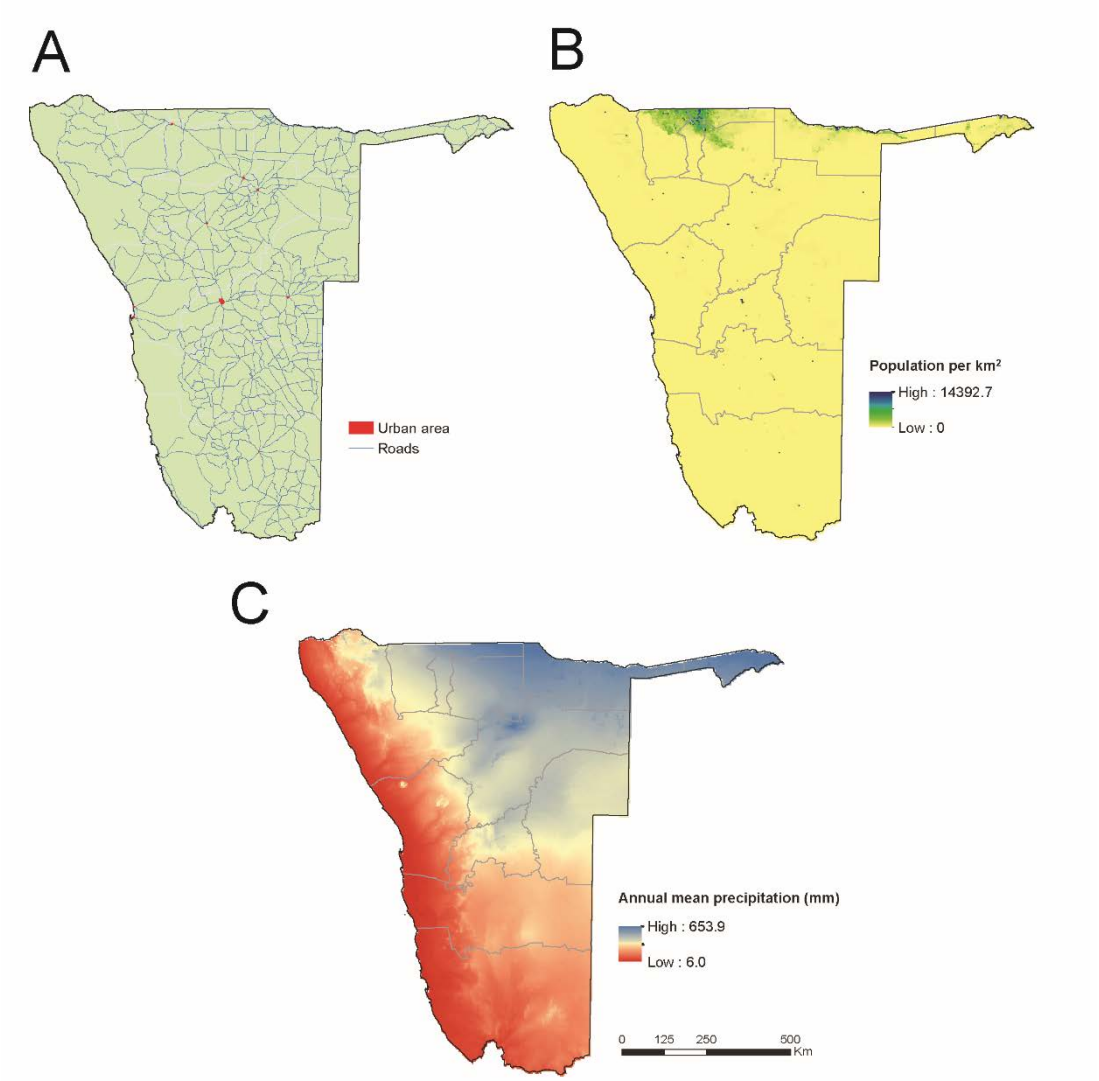

Fig. S1. Urban area and road networks (*A*), population density (*B*), and annual precipitation (*C*) in Namibia. The data on urban areas were obtained from the Natural Earth (www.naturalearthdata.com/downloads/10m-cultural-vectors/10m-urban-area/) (Schneider et al., 2003), and the road networks were obtained from DIVA-GIS (www.diva-gis.org), the population density data were downloaded from WorldPop (www.worldpop.org), and the average annual precipitation for 1970-2000 were obtained from WorldCllim (worldclim.org/version2).

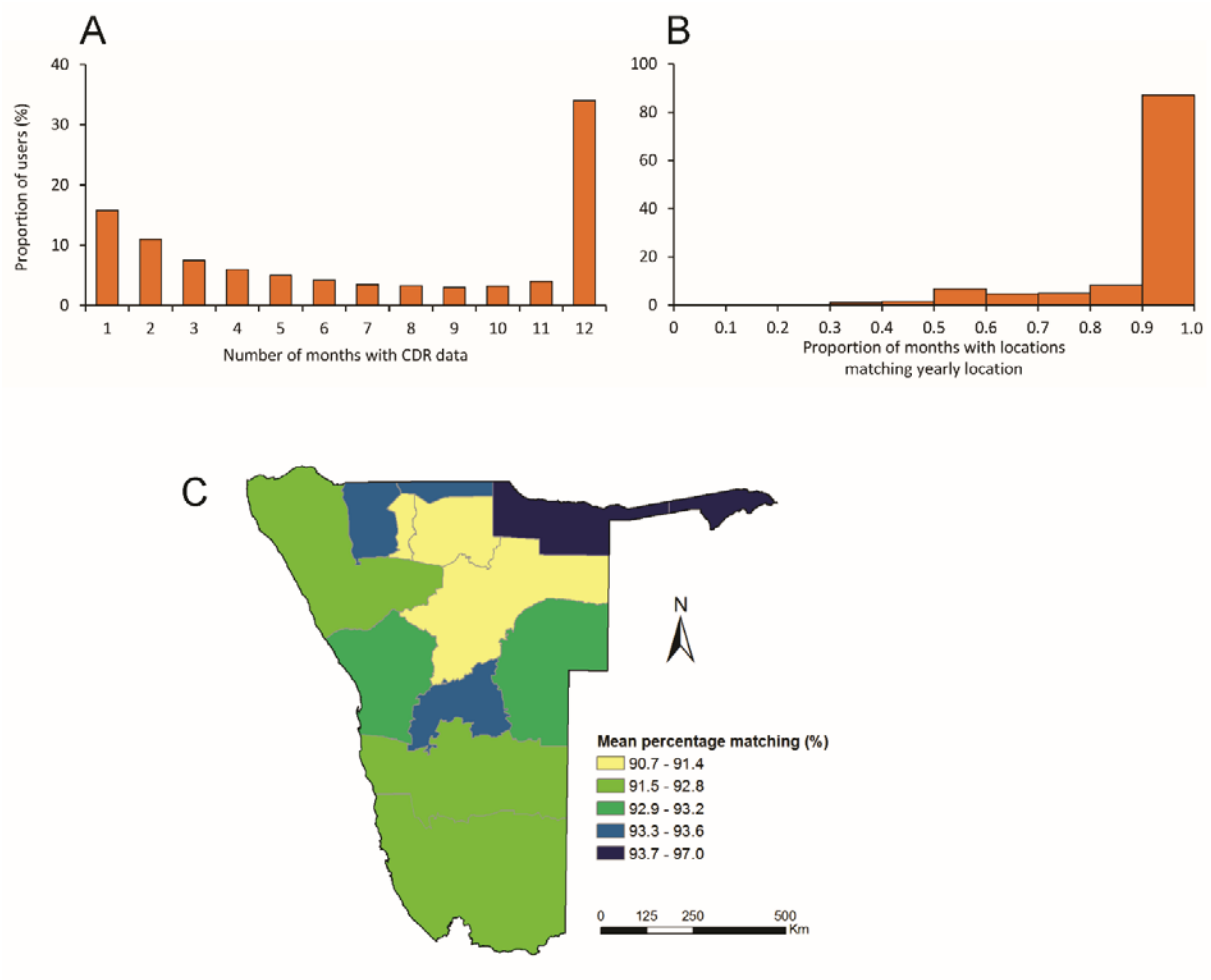

Fig. S2. Comparing CDR-derived monthly and yearly locations of subscribers in 2012. (*A*) Users with monthly CDR data for defining locations. (*B*) Users with monthly locations matching yearly locations, excluding users with only 1-month data. (*C*) Percentage of monthly locations matching yearly locations by region. The monthly/yearly location was defined as the most frequent location at regional level across the whole month/year.

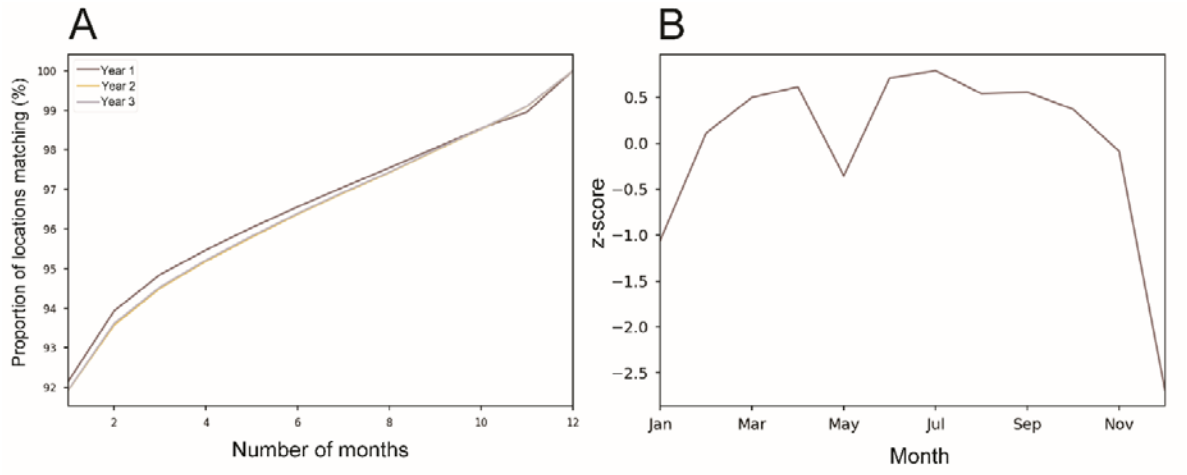

Fig. S3. Comparing the ability of periods of different lengths to define residential location. (*A*) Proportion of monthly locations using different period data and matching the yearly location in three periods: Year 1 (October 2010 – September 2011), Year 2 (October 2011 – September 2012), and Year 3 (October 2012 – September 2013). The proportions in Year 1 and Year 2 are almost identical. (*B*) The Z-score of the percentages of users with monthly locations matching yearly location. The monthly/yearly location was defined as the most frequent location of a phone user over the course of the corresponding period.

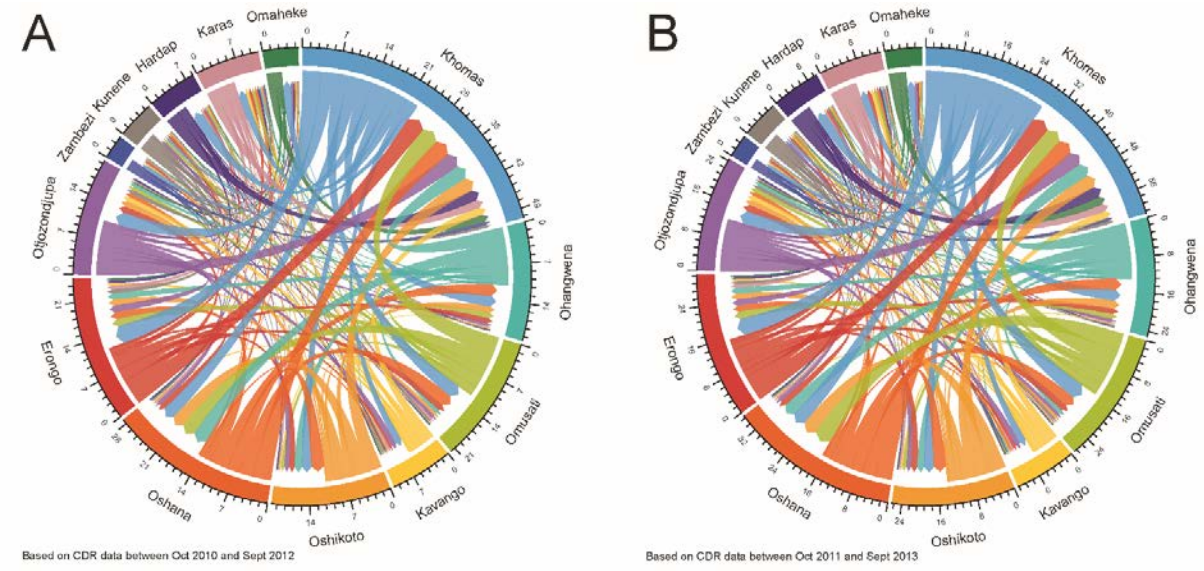

Fig. S4. CDR-derived migrating phone users between regions of Namibia in (*A*) 2011 and (*B*) 2012. The origins and destinations of migrants are each assigned a colour and represented by the circle's segments. The direction of the flow is encoded by both the origin region's colour and a gap between the flow and the destination region's segment. The volume of movement is indicated by the width of flow. Because the flow width is nonlinearly adapted to the curvature, it corresponds to the flow size only at the beginning and end points. Tick marks on the circle segments show the number of migrants (inflows and outflows) in thousands.

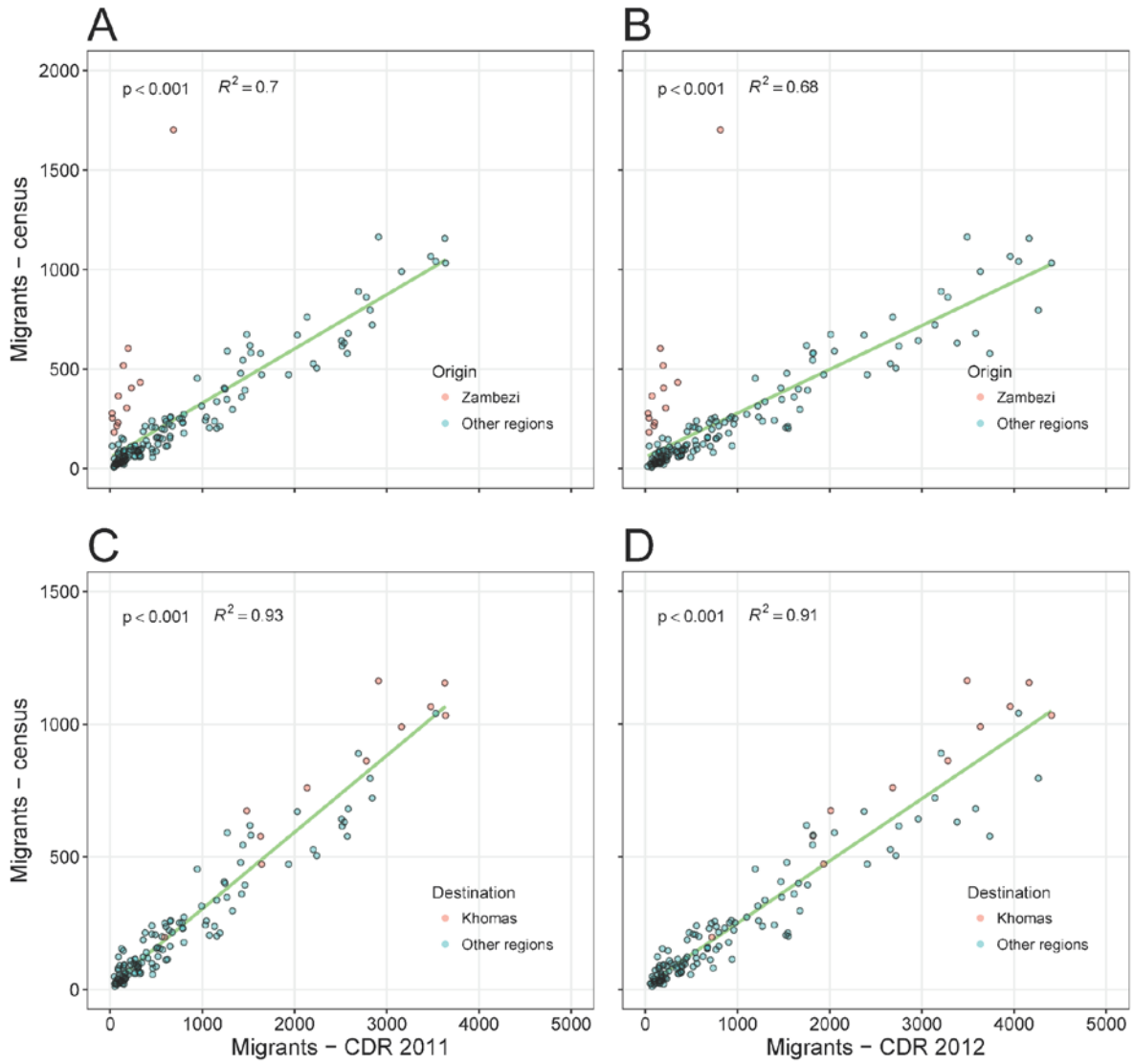

Fig. S5. Relation between census-based migration data and CDR-derived migrating mobile phone users at regional level, with (*A*) and (*B*) for all 13 regions of Namibia, and (*C*) and (*D)* for 12 regions except Zambezi. The green solid lines represent linear regression fit, with *p* values and R-squared (R^2) values presented.

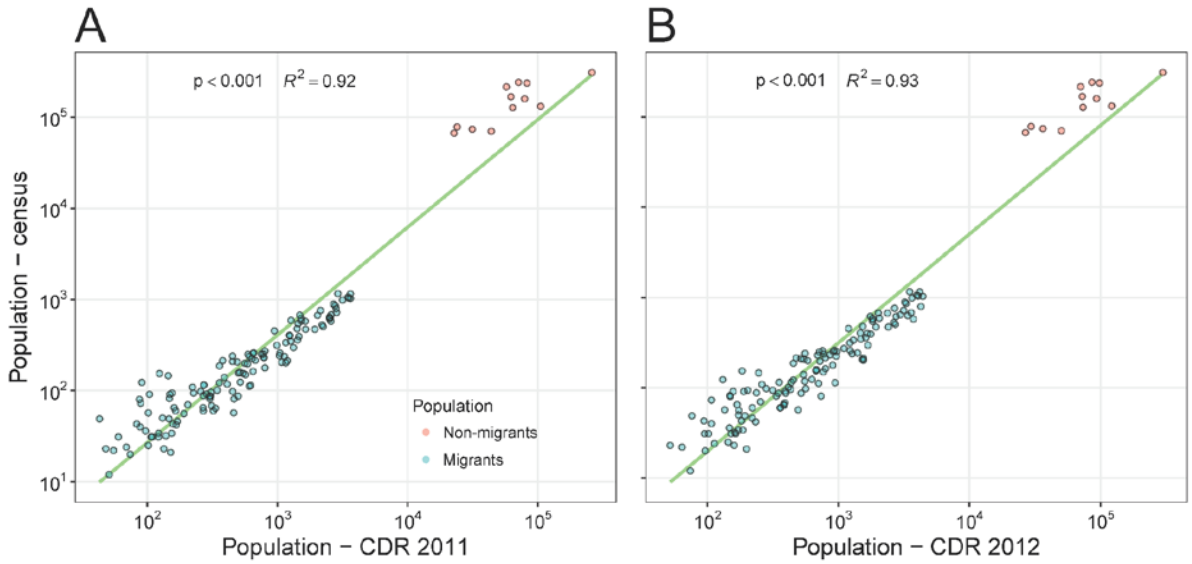

Fig. S6. Relations between 2011 census-derived population and CDR-derived population of mobile phone users in Periods 2011 (*A*) and 2012 (*B*) between regions of Namibia except Zambezi. The blue dots represent the population migrating from one region to another region, and the red dots represent the residents staying in the same region. The green solid lines represent linear regression fit with *p* and R^2 values.

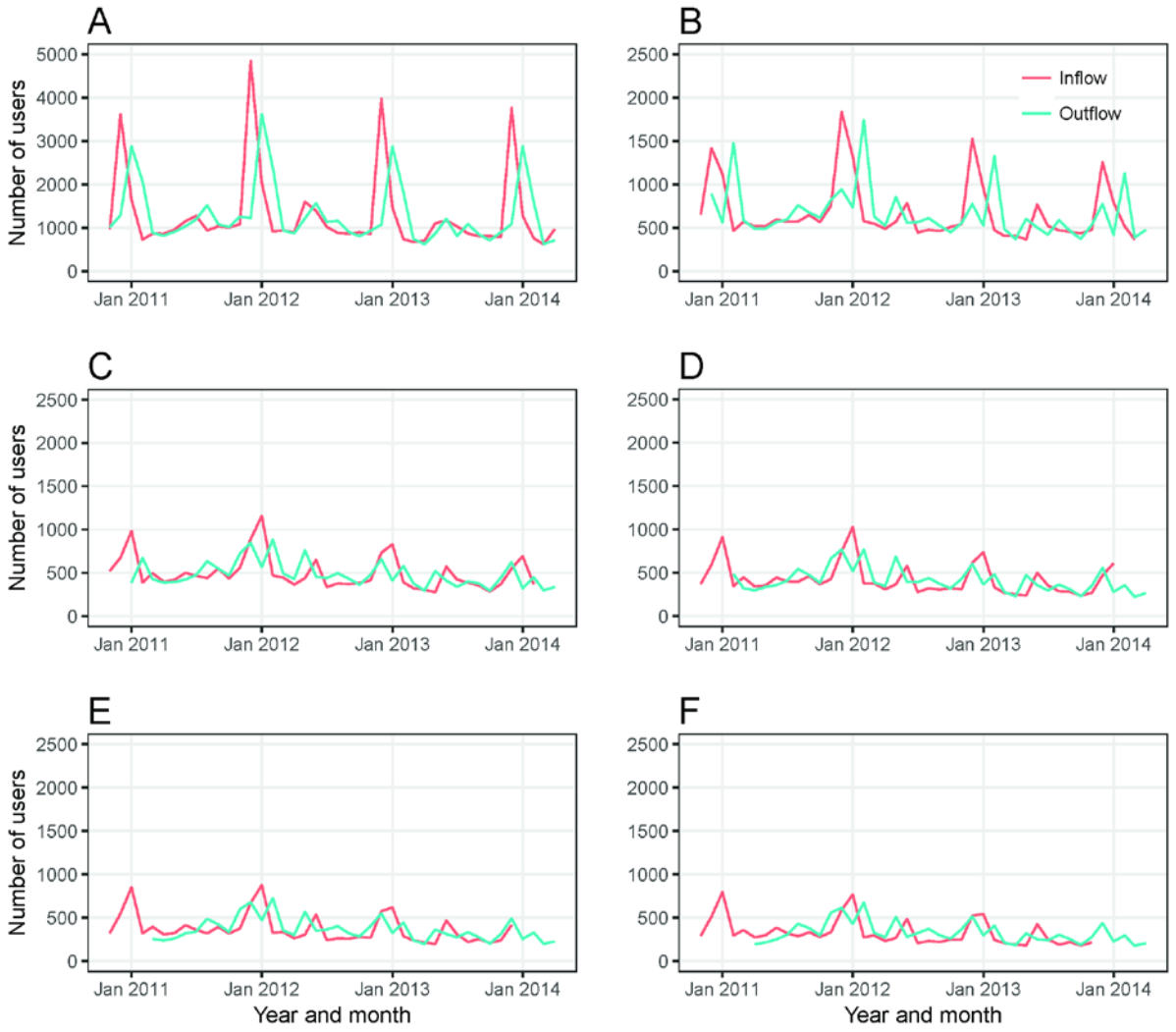

Fig. S7. CDR-derived inflow and outflow of Zambezi with usual residences defined by different lengths of periods, from one month (A) to six months (E). The usual residence at regional level was defined as the most frequent location of a mobile phone user over the course of corresponding period.

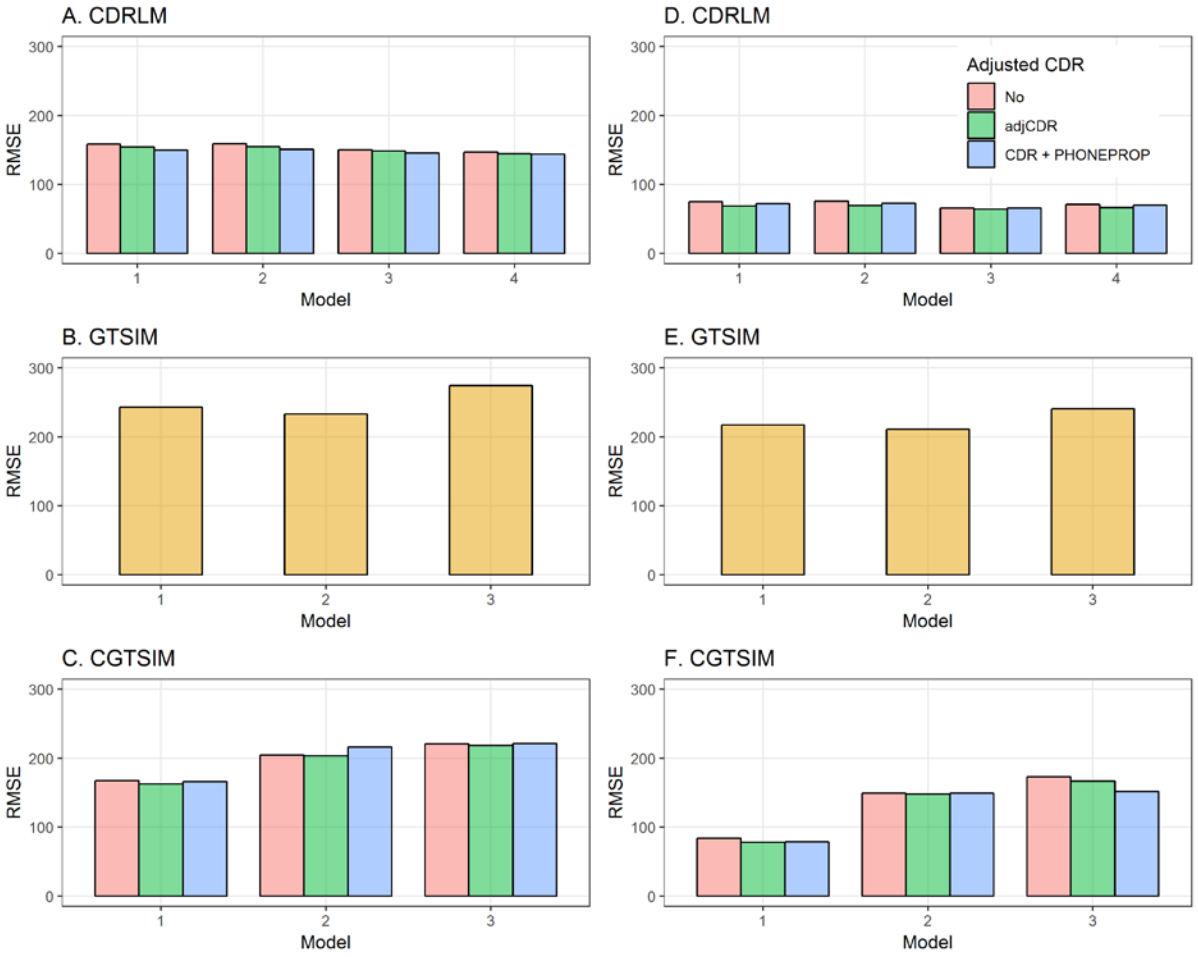

Fig. S8. The root-mean-square error (RMSE) of the CDR-based linear models (CDRLMs), gravity-type spatial interaction models (GTSIMs) and CDR-based gravity-type spatial interaction models (CGTSIMs). The results in (*A*), (*B*), and (*C*) are the RMSE of models tested for all regions, while results in (*D*), (*E*), and (*F*) are the RMSE of models tested for regions except Zambezi. The variables included in these models are detailed in Table S1.

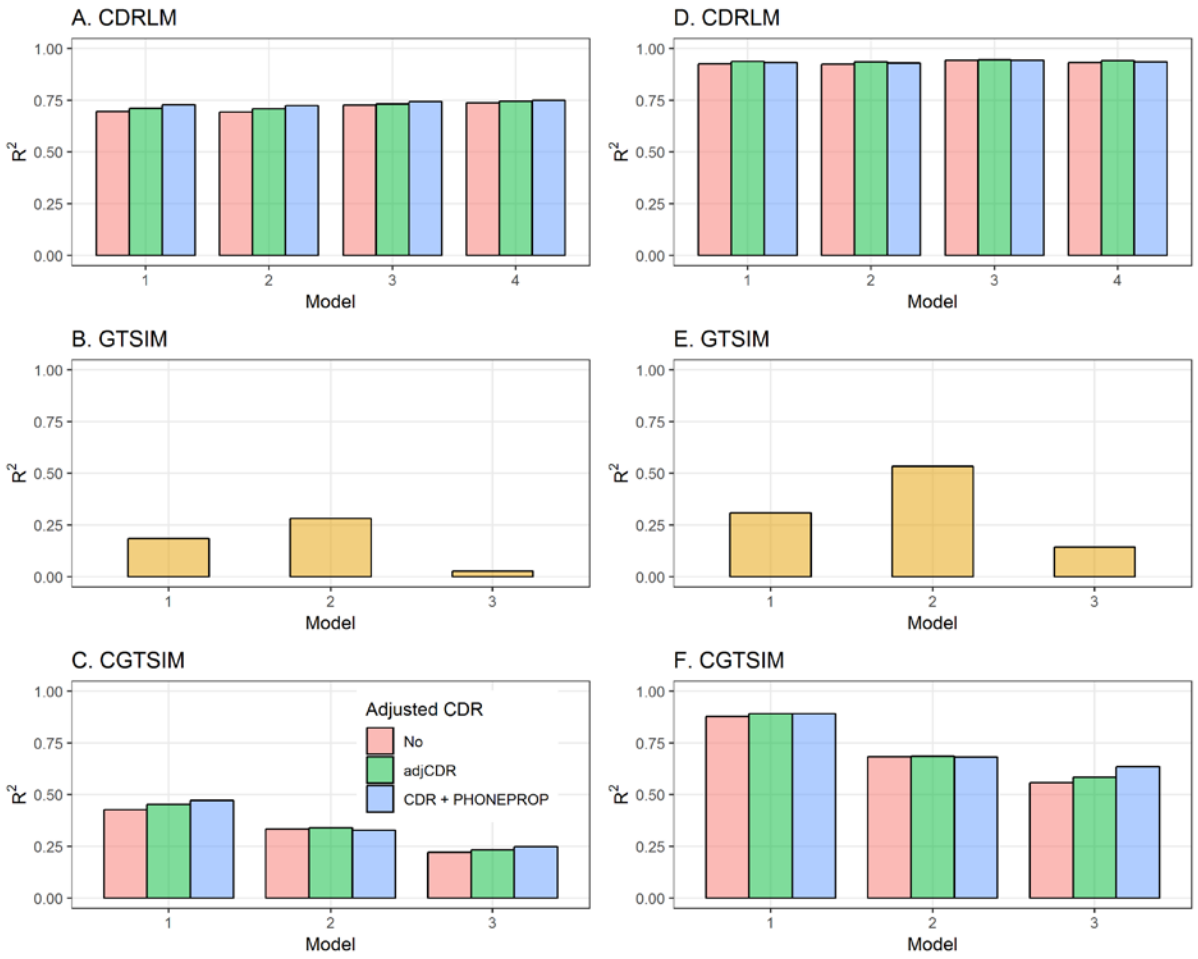

Fig. S9. The R^2 of CDRLMs, GTSIMs and CGTSIMs. The results in (*A*), (*B*), and (*C*) are the R^2 of models tested for all regions, while results in (*D*), (*E*), and (*F*) are the R^2 of models tested for regions except Zambezi. The variables included in these models are detailed in Table S1.

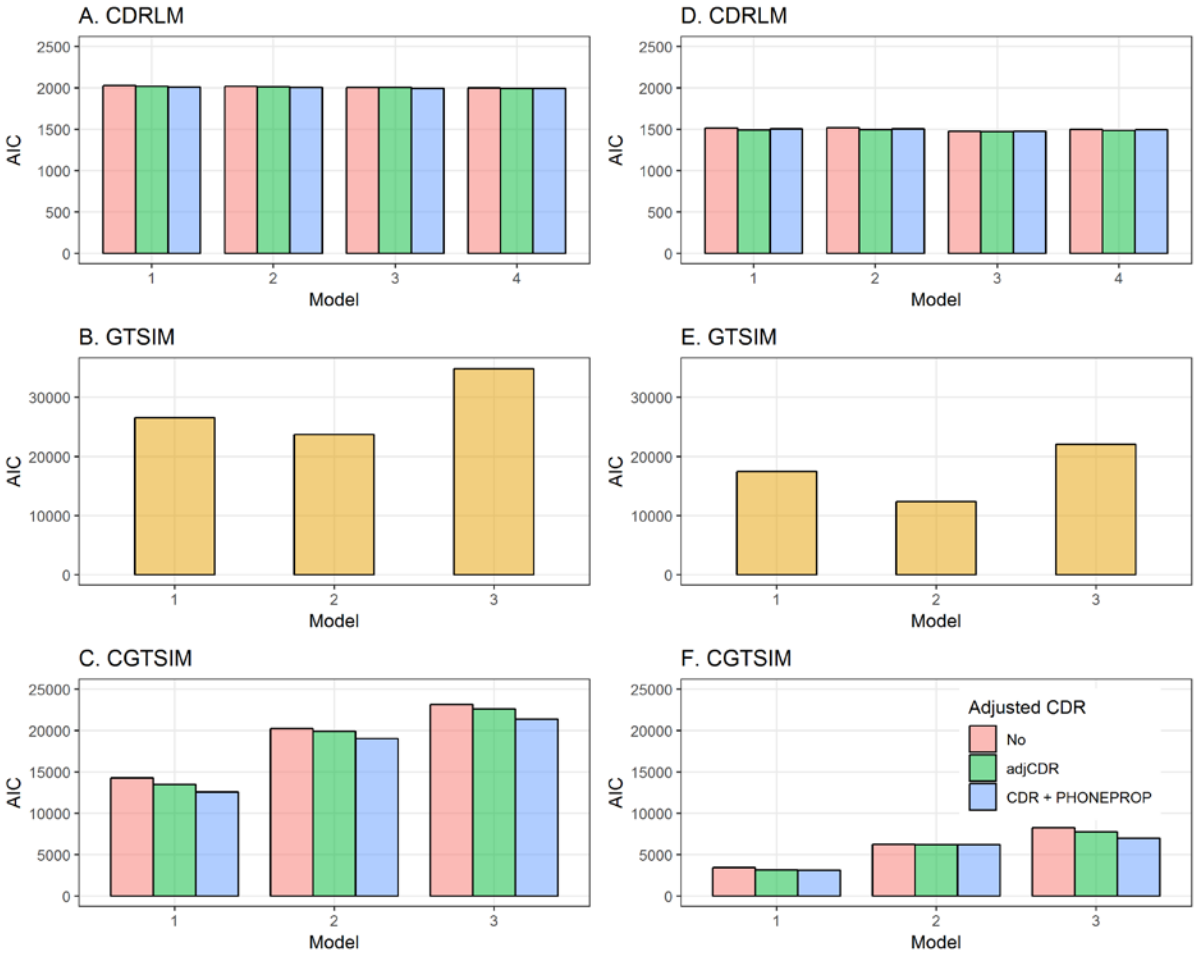

Fig. S10. The Akaike Information Criterion (AIC) for CDRLMs, GTSIMs and CGTSIMs. The results in (*A*), (*B*), and (*C*) are the AIC of models tested for all regions, while results in (*D*), (*E*), and (*F*) are the AIC of models tested for regions except Zambezi. The variables included in these models are detailed in Table S1.

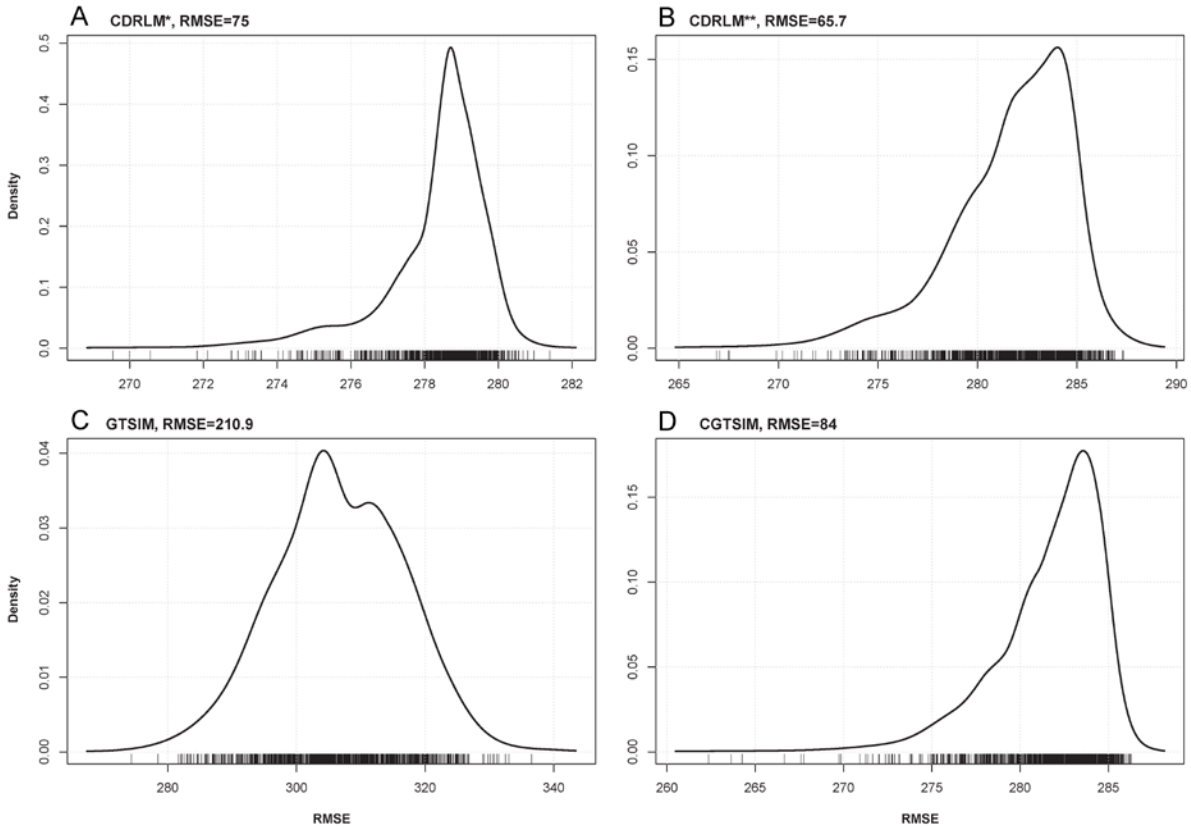

Fig. S11. The distribution of RMSE of models tested by shuffled census-derived migration data through 1000 simulations. (*A*) CDRLM only using a variable of unadjusted CDRs, and (*B*), (*C*) and (*D*) results of optimal CDRLM, GTSIM and CGTSIM, respectively, using unadjusted CDR data and other variables (Fig S8 and Table S1). The RMSE of each model fitted by real, unshuffled census data is given in the title of each graph. The Zambezi region as an outliner is excluded in the dataset.

* The model #1 of CDRLM.

** The model #3 of CDRLM.

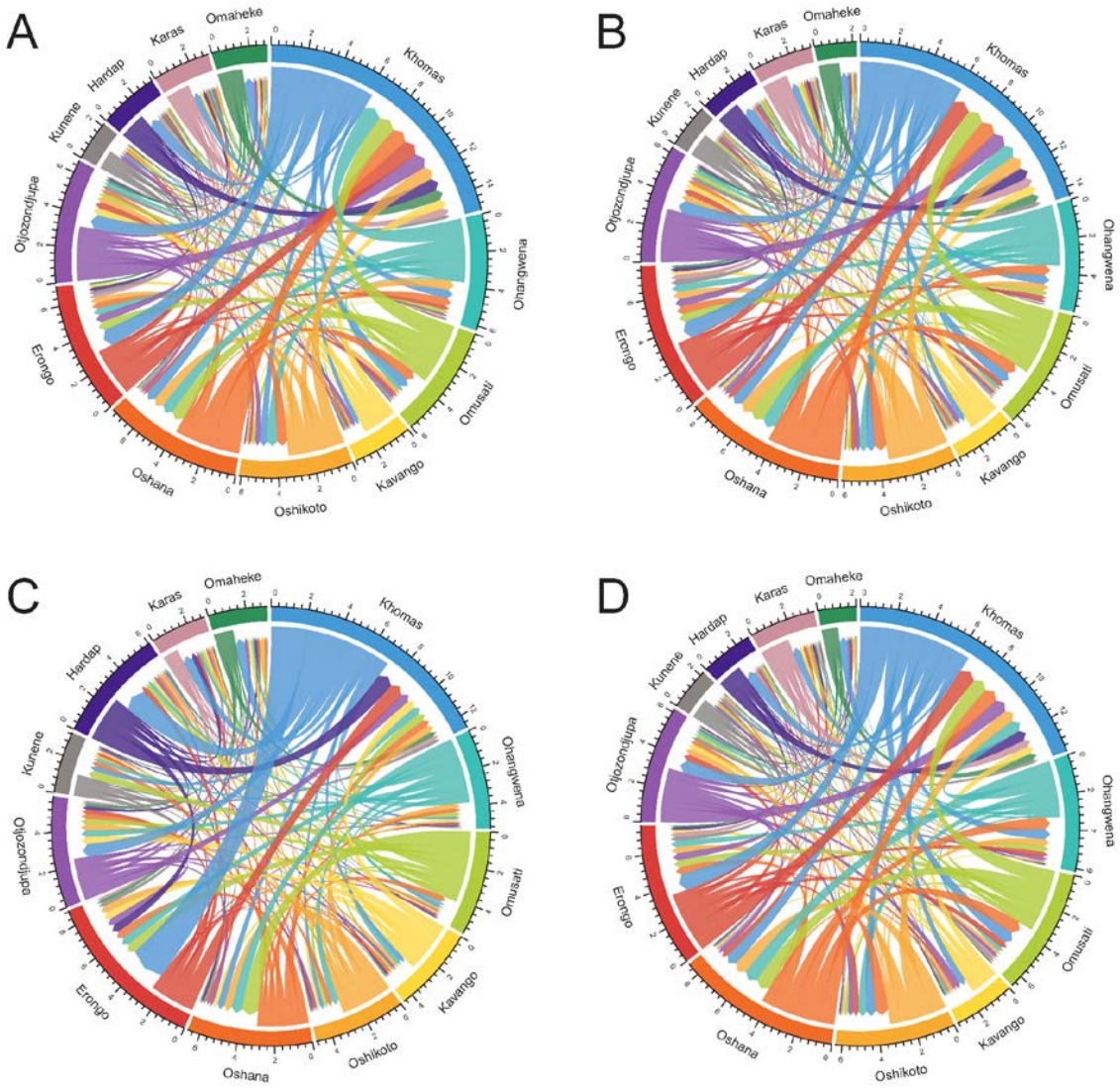

Fig. S12. Comparing census-derived and model-estimated internal migration between regions in Namibia. (*A*) Migration flows derived from the 2011 Namibia Population and Housing Census data. (*B*) Estimates made by the CDRLM. (*C*) Estimates made by GTSIM. (*D*) Estimates made by the CGTSIM. Tick marks on the circle segments show the number of migrants (inflows and outflows) in thousands. The Zambezi region as an outliner is excluded, and the estimates of optimal models with the lowest RMSE using unadjusted CDR data (Fig S8 and Table S1) are presented here.

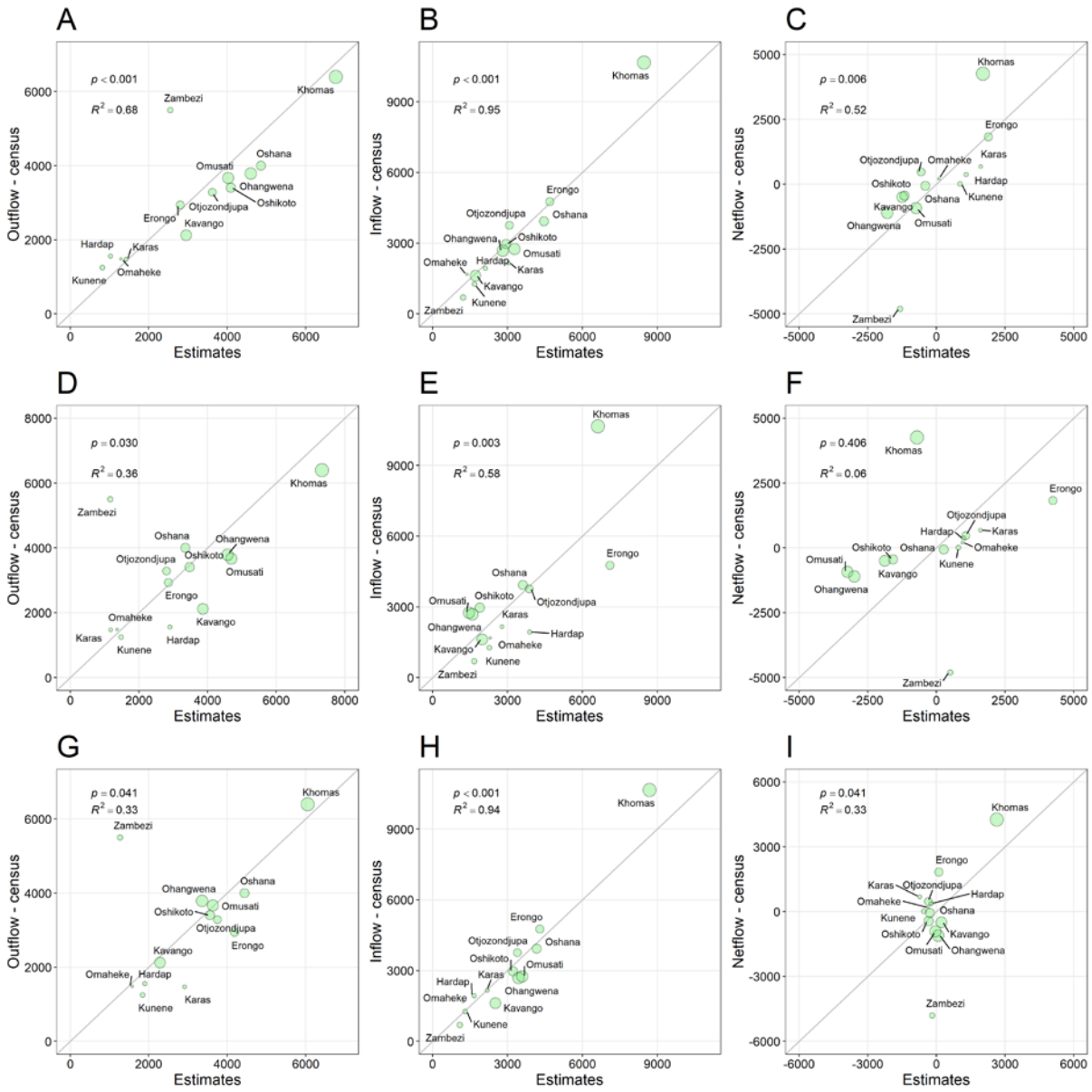

Fig. S13. Comparing regional outflow, inflow and net migration between 2011 census data and estimates made by models for all regions in Namibia. The estimates of optimal models with the lowest RMSE using unadjusted CDR data (Fig S8 and Table S1) are presented, with CDRLM in (*A*), (*B*) and (*C*), GTSIM in (*D*), (*E*) and (*F*), and CGTSIM in (*G*), (*H*) and (*I*).

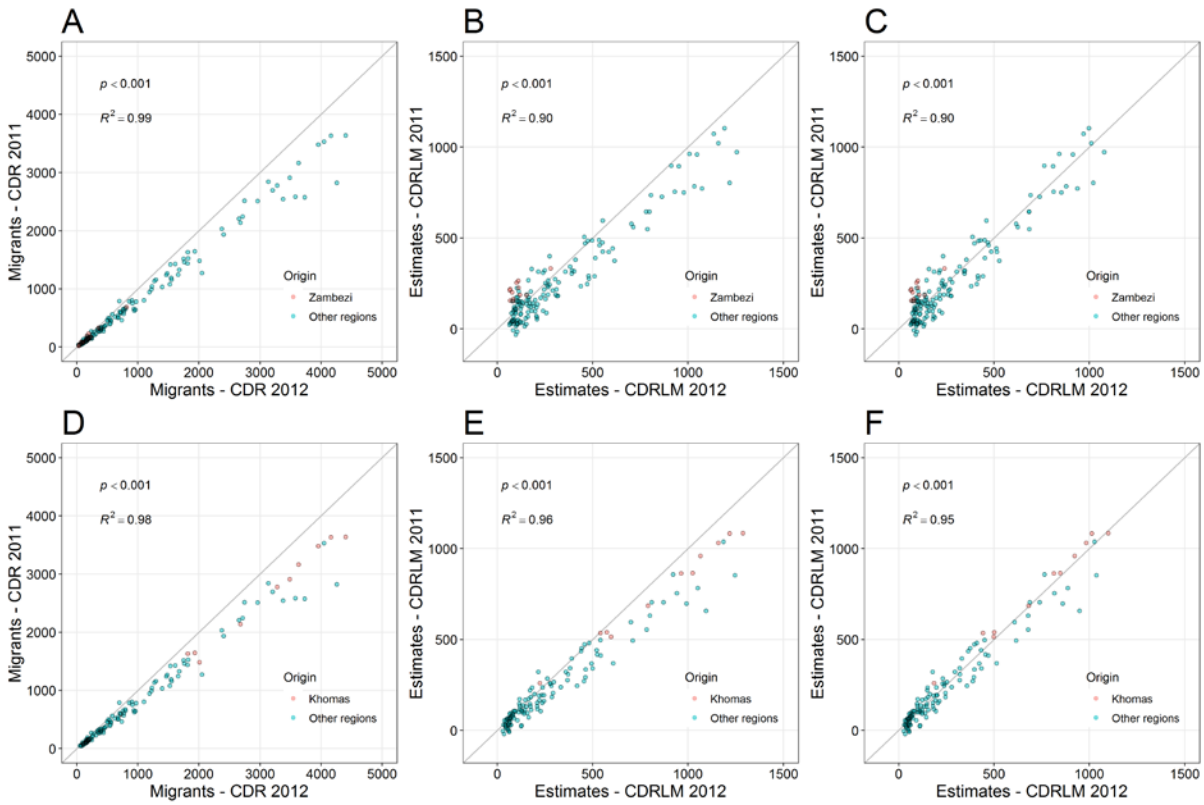

Fig. S14. CDR-derived migrating users and estimates by CDRLM for Periods 2011 and 2012. The graphs in (*A*), (*B*) and (*C*) were plotted for all regions, while Zambezi as an outliner was excluded in (*D*), (*E*) and (*F*). CDR-derived number of migrating mobile phone users are presented in (*A*) and (*D*), respectively. Estimates made by CDRLM are showed in (*B*) and (*E*), and results of CDRLM using CDR data adjusted for the effect of increasing users across years are presented in (*C*) and (*F*). The fitted CDRLMs using only CDRs for 2011 were used to predict migration in 2012.

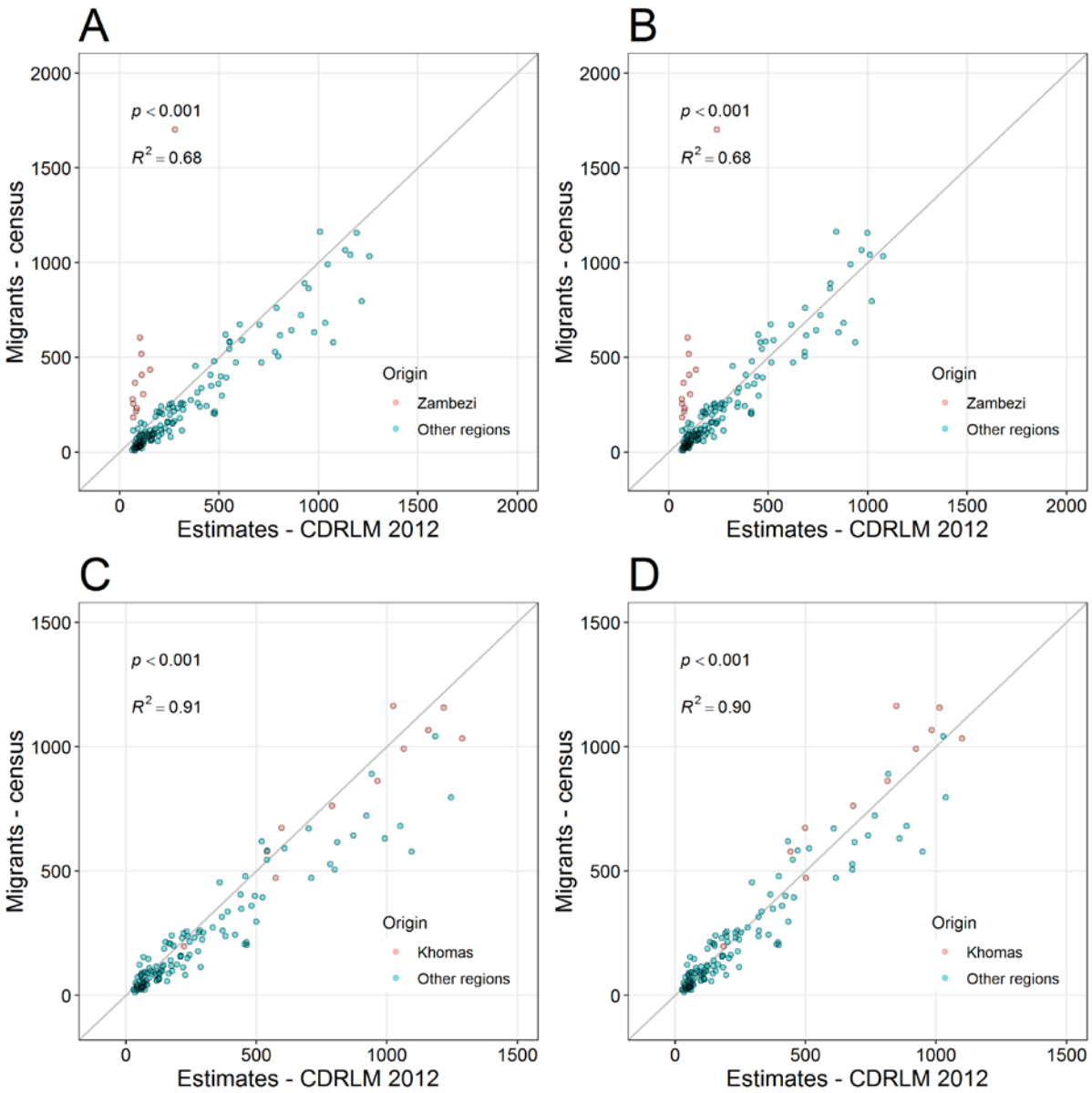

Fig. S15. Comparing census-derived migration in 2011 and estimates made by the CDRLM for 2012. (*A*) Estimates using unadjusted 2012 CDR data for all regions. (*B*) Estimates made using 2012 CDR-derived migrating user data adjusted for the effect of increasing users across years. (*C*) Estimates using unadjusted 2012 CDR data for regions excluding Zambezi. (*D*) Estimates using 2012 CDR data adjusted for the effect of the increasing users across years in regions excluding Zambezi. The fitted CDRLMs using only CDRs for 2011 were used to predict the migration in 2012 with corresponding CDR data.

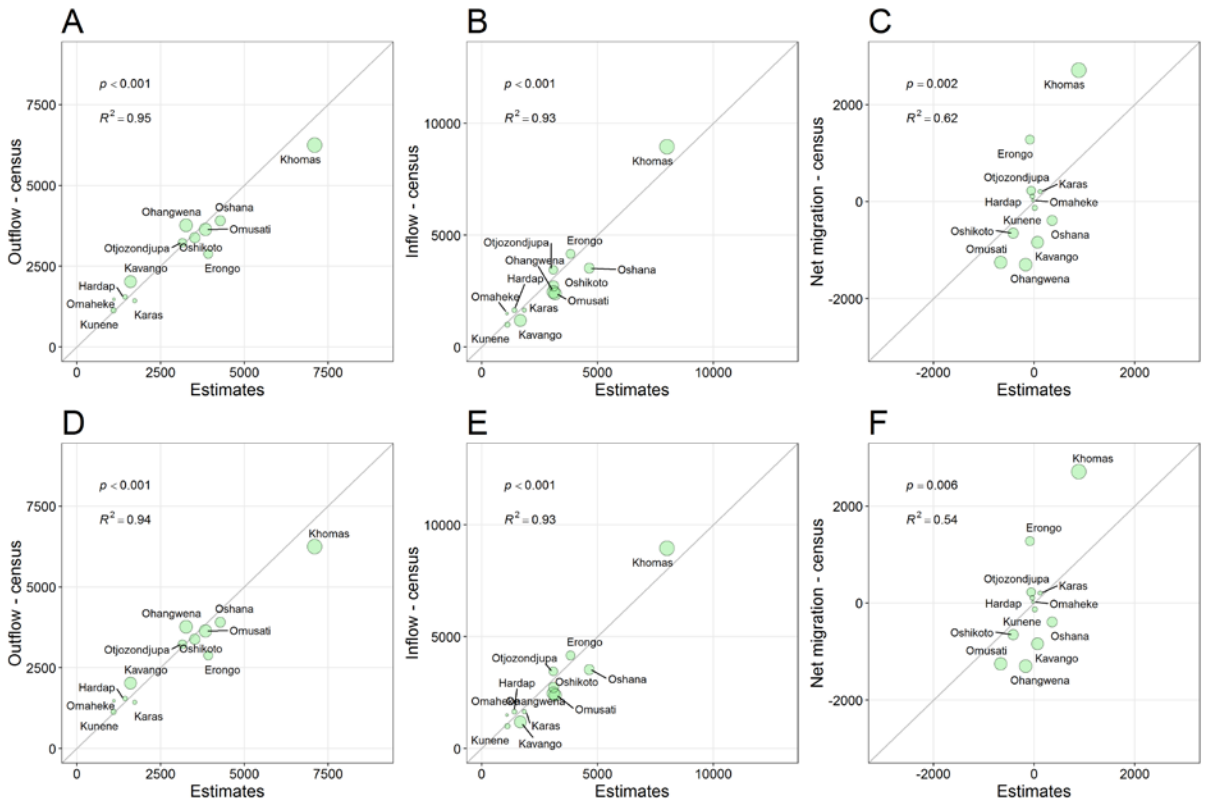

Fig. S16. Comparing regional outflows, inflows and net migration between census data in 2011 and estimates made by CDRLM in 2012. The predictions of CDRLM using unadjusted 2012 CDR data are presented in (*A*), (*B*) and (*C*), while (*D*), (*E*) and (*F*) show the results of CDRLM using CDR data adjusted to offset the effect of increasing number of users across years. The statistics of migration were obtained from the 2011 Namibia Population and Housing Census data. The Zambezi region as an outliner is excluded in these graphs, and the fitted CDRLMs using only CDRs for 2011 were used to predict the migration in 2012 with corresponding CDR data.

Reference

Callegaro M and Poggio T (2004) Where can I call you? The "Mobile (Phone) Revolution" and its impact on survey research and coverage error: a discussion of the Italian case.

GADM (2018) *Database of Global Administrative Areas (version 3.4)*. Available from: http://www.gadm.org [Accessed 24 April 2018]

Garcia A J, Pindolia D K, Lopiano K K and Tatem A J (2015) Modeling internal migration flows in sub-Saharan Africa using census microdata. *Migration Studies*; **3**(1): 89-110.

Henry S, Boyle P and Lambin E F (2003) Modelling inter-provincial migration in Burkina Faso, West Africa: the role of socio-demographic and environmental factors. *Applied Geography*; **23**(2-3): 115-136.

Henry S, Piche V, Ouedraogo D and Lambin E F (2004) Descriptive analysis of the individual migratory pathways according to environmental typologies. *Population and Environment*; **25**(5): 397-422.

Lall S V, Selod H and World Bank. (2006) *Rural-urban migration in developing countries a survey of theoretical predictions and empirical findings*. Washington, D.C.: World Bank, Available from: http://econ.worldbank.org/external/default/main?pagePK=64165259&theSitePK=469382&piPK=64165421&menuPK=64166093&entityID=000016406_20060505110833

Mobile Telecommunications (2012) *MTC Annual Report 2012*. Available from: http://www.mtc.com.na/sites/annual-reports/2012/Annual_Report_2012.pdf [Accessed 21 November 2017]

Mobile Telecommunications (2014) *MTC Annual Report 2014*. Available from: http://www.mtc.com.na/sites/annual-reports/2014/pdf/MTCAnnualReport2014.pdf [Accessed 23 November 2017]

Mobile Telecommunications (2018) *Coverage*. Available from: http://www.mtc.com.na/coverage [Accessed 3 June 2018]

Namibia Statistics Agency (2013) *Namibia 2011 Population & Housing Census Main Report*. Available from: https://cms.my.na/assets/documents/p19dmn58guram30ttun89rdrp1.pdf [Accessed 22 November 2017]

Ruktanonchai N W, Bhavnani D, Sorichetta A, Bengtsson L, Carter K H, Cordoba R C, Le Menach A, Lu X, Wetter E, Erbach-Schoenberg E Z and Tatem A J (2016) Census-derived migration data as a tool for informing malaria elimination policy. *Malaria Journal*; **15**

Schneider A, Friedl M A, McIver D K and Woodcock C E (2003) Mapping Urban Areas by Fusing Multiple Sources of Coarse Resolution Remotely Sensed Data. *Photogrammetric Engineering & Remote Sensing*; **69**(12): 1377-1386.

Smits J and Steendijk R (2015) The international wealth index (IWI). *Social Indicators Research*; **122**(1): 65-85.

Sorichetta A, Bird T J, Ruktanonchai N W, Erbach-Schoenberg E Z, Pezzulo C, Tejedor N, Waldock I C, Sadler J D, Garcia A J, Sedda L and Tatem A J (2016) Mapping internal connectivity through human migration in malaria endemic countries. *Scientific Data*; **3**

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

Vobruba T, Korner A and Breitenecker F (2016) Modelling, Analysis and Simulation of a Spatial Interaction Model. *Ifac Papersonline*; **49**(29): 221-225.

Wesolowski A, O'Meara W P, Eagle N, Tatem A J and Buckee C O (2015) Evaluating Spatial Interaction Models for Regional Mobility in Sub-Saharan Africa. *Plos Computational Biology*; **11**(7)

Wesolowski A, Zu Erbach-Schoenberg E, Tatem A J, Lourenco C, Viboud C, Charu V, Eagle N, Engo-Monsen K, Qureshi T, Buckee C O and Metcalf C J E (2017) Multinational patterns of seasonal asymmetry in human movement influence infectious disease dynamics. *Nature Communications*; **8**(1): 2069.