Dynamic population mapping using mobile phone data

Pierre Deville[a,b,c,1], Catherine Linard[c,d,1,2], Samuel Martin[e], Marius Gilbert[c,d], Forrest R. Stevens[f], Andrea E. Gaughan[f], Vincent D. Blondel[a], and Andrew J. Tatem[g,h,i]

[a]Department of Applied Mathematics, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium; [b]Center for Complex Network Research and Physics Department, Northeastern University, Boston, MA 02115; [c]Fonds National de la Recherche Scientifique, B-1000 Brussels, Belgium; [d]Biological Control and Spatial Ecology, Université Libre de Bruxelles, B-1050 Brussels, Belgium; [e]Université de Lorraine CNRS, Centre de Recherche en Automatique de Nancy, UMR 7039, 54518 Vandoeuvre-lès-Nancy, France [f]Department of Geography and Geosciences, University of Louisville, Louisville, KY 40292; [g]Department of Geography and Environment, University of Southampton, Southampton SO17 1BJ, United Kingdom; [h]Fogarty International Center, National Institutes of Health, Bethesda, MD 20892; and [i]Flowminder Foundation, 17177 Stockholm, Sweden

Edited by Michael F. Goodchild, University of California, Santa Barbara, CA, and approved September 15, 2014 (received for review May 8, 2014)

During the past few decades, technologies such as remote sensing, geographical information systems, and global positioning systems have transformed the way the distribution of human population is studied and modeled in space and time. However, the mapping of populations remains constrained by the logistics of censuses and surveys. Consequently, spatially detailed changes across scales of days, weeks, or months, or even year to year, are difficult to assess and limit the application of human population maps in situations in which timely information is required, such as disasters, conflicts, or epidemics. Mobile phones (MPs) now have an extremely high penetration rate across the globe, and analyzing the spatiotemporal distribution of MP calls geolocated to the tower level may overcome many limitations of census-based approaches, provided that the use of MP data is properly assessed and calibrated. Using datasets of more than 1 billion MP call records from Portugal and France, we show how spatially and temporarily explicit estimations of population densities can be produced at national scales, and how these estimates compare with outputs produced using alternative human population mapping methods. We also demonstrate how maps of human population changes can be produced over multiple timescales while preserving the anonymity of MP users. With similar data being collected every day by MP network providers across the world, the prospect of being able to map contemporary and changing human population distributions over relatively short intervals exists, paving the way for new applications and a near real-time understanding of patterns and processes in human geography.

population distribution | phone calls | human mobility | census | remote sensing

Our knowledge of human population numbers and distribution for many areas of the world remains poor (1) despite their importance for policy (2, 3), operational decisions (4), and research (5–7) across many fields. In the 1990s, a growing interest in the global mapping of human populations emerged (8, 9), leading to the advanced development of methodologies that undertake the spatial downscaling of human population count data from censuses summarized over large and irregular administrative units to grid squares of 100 m to 5 km resolution (10–16). Initial efforts to downscale these data used simple areal weighting methods (10, 17) or dasymetric modeling approaches (13–15), which use ancillary layers to redistribute population counts within administrative units (18). Modeling techniques that spatially downscale population numbers into gridded datasets continue to be refined, with basic dasymetric models increasing in sophistication, incorporating multiscale remotely sensed and geospatial data and making improvements in the type of statistical algorithms used in the modeling process (19–21). These detailed population databases have proven crucial for studies reliant on information about human population distributions, typically for calculating populations at risk for human or natural disasters (22–24), to assess vulnerabilities (7, 25), or to derive health and development indicators (3, 5, 26, 27). However, despite improvements, these data still have many limitations.

Regardless of how sophisticated these methods are, they remain largely constrained by population count data from censuses that form the basis for the estimation of population distributions across large areas (10–17). Although the increasing use of global positioning and geographical information system technologies has supported the improved collection of census data and their processing, censuses remain an infrequent and expensive source of detailed population data. Moreover, for many low-income countries, the unreliability of estimates, low spatial resolution, and complete lack of contemporary data represent further limitations. These restrictions mean that the latest health indicators or estimates of populations at risk often may be based on outdated and coarse input population data (26, 28, 29), a particularly restrictive feature when accurate contemporary numbers may be required for disaster impact assessments, epidemic modeling, or conflict relief planning. Human populations are dynamic, moving daily, seasonally, and annually, resulting in rapidly changing densities. Attempts have been made to model and map these dynamics for high-income countries (20, 30), but the data streams upon which such models are based currently are unavailable to most of the world, particularly resource-poor regions.

The proliferation of mobile phones (MPs) offers an unprecedented solution to this data gap. The global MP penetration

Significance

Knowing where people are is critical for accurate impact assessments and intervention planning, particularly those focused on population health, food security, climate change, conflicts, and natural disasters. This study demonstrates how data collected by mobile phone network operators can cost-effectively provide accurate and detailed maps of population distribution over national scales and any time period while guaranteeing phone users' privacy. The methods outlined may be applied to estimate human population densities in low-income countries where data on population distributions may be scarce, outdated, and unreliable, or to estimate temporal variations in population density. The work highlights how facilitating access to anonymized mobile phone data might enable fast and cheap production of population maps in emergency and data-scarce situations.

Author contributions: P.D., C.L., S.M., M.G., V.D.B., and A.J.T. designed research; P.D. and C.L. performed research; F.R.S. and A.E.G. contributed new reagents/analytic tools; P.D., C.L., and S.M. analyzed data; and P.D., C.L., M.G., and A.J.T. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]P.D. and C.L. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: linard.catherine@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1408439111/-/DCSupplemental.

rate (i.e., the percentage of active MP subscriptions within the population) reached 96% in 2014 (31). In developed countries, the number of MP subscribers has surpassed the total population, with a penetration rate now reaching 121%, whereas in developing countries, it is as high as 90% and continuing to rise (31). MP networks, also called cellular networks, are composed of cells, i.e., geographic zones around a phone tower. Each MP communication can be located by identifying the geographic coordinates of its transmitting tower and the associated cell. This network-based positioning method is simple to implement, and its accuracy depends directly upon the network structure; the higher the density of towers, the higher the precision of the MP communication geolocalization (32). Records detailing the time and associated cell of calls and text messages from anonymous users therefore provide a valuable indicator of human presence, and coupled with the increasing use of MPs, offer a promising alternative data source for increasing the spatial and temporal detail of large-scale population datasets. Data provided by communication tools are opening up new opportunities for studying sociospatial behaviors (33–36). MP call detail records were used in the past for studying human mobility patterns at the individual level (37–39) or for mapping human movements and activities using aggregated data (40–44). Most of these studies focused on specific cities or city neighborhoods or groups, and were aimed at understanding traffic flows (40), mapping the intensity of human activities at different times (42–44), or exploring seasonality in foreign tourist numbers and destinations (45, 46). Population movement analyses based on MP data are particularly promising for improving responses to disasters (47, 48) and for planning malaria elimination strategies (49–51). However, to date, these data have not been assessed in their capacity to map human population at fine spatial and temporal resolutions over large geographical extents.

Using Portugal and France as case studies, this study examines how aggregated MP data might be used efficiently to map population distributions at the country scale and reveal otherwise unmeasurable patterns in space and time. We also assess how such predictions compare with existing state-of-the-art downscaling methods. To facilitate widespread use, the methodologies were designed to be easy to implement while minimizing the impact of phone use and network coverage heterogeneities across social groups, regions, and network providers.

Results

The ability of the MP data-based approach to accurately downscale census population data was compared with that of an existing method used to downscale census data through remote sensing and other geospatial data (19), hereafter called the "remote sensing" method or RS (*SI Appendix*, section A.1). Fig. 1 shows the nighttime maps produced for Portugal using the MP (Fig. 1 *B* and *E*) and RS methods (Fig. 1 *C* and *F*), compared with baseline census-derived population densities (Fig. 1 *A* and *D*). At the national scale, both methods show similar spatial patterns that match baseline data, with major cities being clearly identifiable (Fig. 1 *A*–*C*). However, the close-up on the capital city of Lisbon highlights clear differences in estimated population densities visible at finer spatial scales (Fig. 1 *D*–*F*). The spatial detail of the MP method relies on the density of towers, which is substantially higher in urban areas, whereas the spatial detail of the RS method depends on the spatial resolution of the geospatial datasets used in the mapping process, which often do not capture intraurban variations.

Precision and accuracy statistics, including the Pearson product–moment correlation coefficient (r) and root-mean-square error (rmse) were calculated to compare the performance of the MP and RS downscaling methods, using the baseline census-derived population densities as a reference (Fig. 2). The wider cloud observed for the MP method (Fig. 2*A*) indicates a lower precision, especially in low-density areas. The RS method

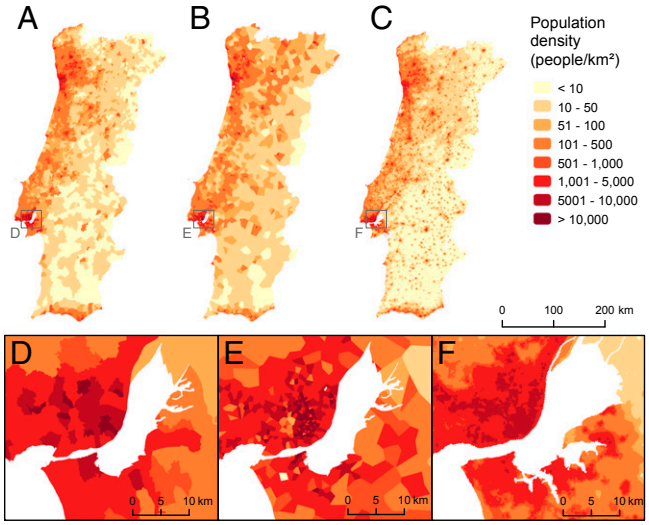

Fig. 1. Comparison of predicted population density datasets with baseline data for mainland Portugal. (*A*) Population density as calculated from the national census at administrative unit level 5 (ADM-5; freguesia). (*B*) Population density at the level of Voronoi polygons, as estimated by the MP method. (*C*) Population density at the level of 100 × 100-m grid squares, as estimated by the RS method. (*D*–*F*) Close-ups around the capital city Lisbon.

produced a higher precision but less accurate predictions, with an overestimation of population densities in low-density areas and an underestimation of population densities in high-density areas (Fig. 2*B*). Globally, the RS method was found to be more precise than the MP method ($r^{MP} = 0.89$; $r^{RS} = 0.92$). Fig. 2*C* shows how the normalized rmse of both methods decreases with population density. A similar but inverse trend was observed for r, with a general increase of r values with population density. Rmse values were always higher for the MP than the RS method, except in high-density areas. Overall, however, the MP method was found to be slightly more accurate than the RS method ($\text{rmse}^{MP} = 796$; $\text{rmse}^{RS} = 850$), given the importance of densely populated areas in the rmse calculation. As shown in *SI Appendix*, section A.3, a combination of both methods further improved the accuracy of the population mapping, highlighting the complementarity of the two approaches.

To assess the robustness of the MP downscaling method and its extrapolation ability, we quantified the impact of the choice of training data on parameter estimations and analyzed the variability of parameter estimations within (*SI Appendix*, section B) and between countries (*SI Appendix*, section C.4). The population density (ρ_c) in a given area c was estimated as a function of the nighttime MP user density (σ_c) for that area by $\rho_c = \alpha\sigma_c^{\beta}$, where the parameters α and β were fitted by a linear regression based on training data. The parameter α represents the ratio between MP user density and population density, which is adjusted by using the census-derived national population. The parameter β reflects the superlinear effect of densely populated areas on human activities. In previously published studies, β was reported to be slightly below 1 and to show little variation (52–55). Although these previously published estimates were obtained based on the number of calls or users per MP tower, rather than on the density of calls or users in a tower's covering area, similar values were expected in our analysis.

By using a standard cross-validation procedure in Portugal, best-fit estimates of 62.95 ± 2.48 for α and 0.803 ± 0.015 for β were found, whereas these estimations became 69.11 ± 10.49 for α and 0.767 ± 0.055 for β when using a spatially stratified cross-validation procedure (*SI Appendix*, section B.2). Such a spatially stratified cross-validation procedure, in which training and test

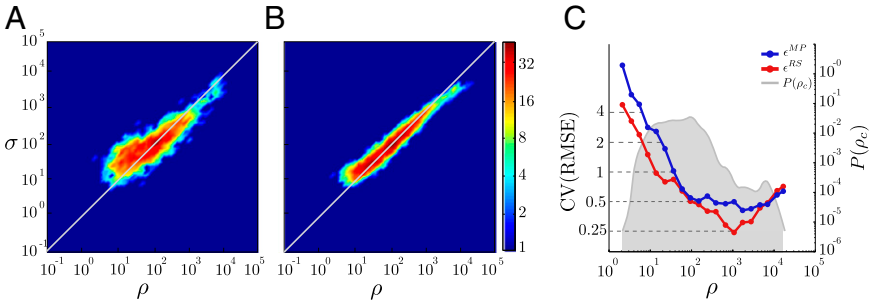

Fig. 2. Precision and accuracy assessments of the MP and RS methods in Portugal. Relation between baseline and estimated population densities using (*A*) the MP method and (*B*) the RS method. (*C*) Rmses normalized by the average population density of intervals for the MP (blue) and RS (red) methods on a logarithmic scale. The shaded area represents the absolute population count per interval. Both methods were calibrated on the Norte region (n = 1,425), and their accuracy was assessed on the rest of the country (n = 1,457).

sets are sampled from geographically distinct regions (56), allowed for a quantitative assessment of the extrapolation capacity of the model (57, 58). Here, the larger confidence intervals obtained using the spatially stratified cross-validation procedure reflect the impact of spatially clustered population densities on the estimation of SEs. This variability is important to take into account when extrapolating the model to a data-scarce and geographically different region. The accuracy and precision of population density estimates are not sensitive to the estimate of α, as changes in α values are corrected by total population adjustments. However, results showed a relatively high sensitivity to the estimate of β, with an rmse increase of up to 15% for β values within the larger confidence interval (0.77 ± 0.055) (*SI Appendix*, section B.3). Here, however, β was found to be relatively stable both within and between countries, the best-fit estimates being 0.902 ± 0.036 and 0.846 ± 0.056 in France, using the standard and spatially stratified cross-validation procedures, respectively (*SI Appendix*, section C.4).

To be widely applied and to facilitate the acquisition of MP data, the method may be simplified by using the density of phone calls instead of the density of different users over a certain time window. This was done for data from France, where information on users was not accessible. Even if the resulting population density datasets were slightly less accurate—although not always significantly—the very similar estimated β values (*SI Appendix*, section C.2) and the very low spatiotemporal variations in MP use behaviors (*SI Appendix*, section C.3) suggest a minimal effect on population density estimates. Similarly, daily-aggregated MP data may be used instead of nighttime data when the time of MP calls is not known, although that may induce higher uncertainty in population density estimates as the model is calibrated using census-derived nighttime data. However, the precise accuracy loss cannot be estimated here, because daytime data would be required as a reference for accuracy assessment (*SI Appendix*, section C.2).

The potential of MP data to estimate population density variations through time is illustrated in Fig. 3. The relative differences in estimated population densities between the major holiday period (July and August) and more traditional working periods (from September to June) in Portugal and France reveal clear spatial patterns (Fig. 3). Seasonal changes in population distribution are evident: most cities are characterized by a large decrease in population densities during the holiday period, whereas less-populated areas and well-known tourist sites, such as coastlines or mountainous areas, show large increases. Fig. 3*E* shows that population densities decrease in Paris, with the exception of a few spots corresponding to highly visited sites (e.g., Disneyland Paris, Charles de Gaulle airport). Maps of daily and weekly population dynamics in Portugal and France are shown in *SI Appendix*, section D. In addition to providing quantitative measures of how people from densely populated areas tend to travel toward more low-density and recreational locations during holidays or weekends, this method also offers a detailed visualization and quantification of the dynamic popularity of a given place over time.

Discussion

The increasing penetration of mobile phones and other information and communication tools used daily by a large proportion of the global population offers a wealth of new spatiotemporal data that are contributing to the "big data" revolution. These new data have the potential to profoundly transform the way we think about and conduct science, especially geographical analyses, as most of these data are implicitly or explicitly spatial (59, 60). In operational and governmental decisions, these data also may be valuable for supporting rapid responses to disruptive events or longer-term planning purposes. In the specific application presented here, spatially and temporally detailed population distribution datasets potentially may provide the essential denominator required in many fields, such as studying collective human responses to disease outbreaks (61, 62), emergencies (63, 64), or any application for which information on daily, seasonal, or annual changes in population distribution is useful.

This study demonstrates how the analysis of MP data that are collected readily every day by phone network providers can complement traditional census outputs. Not only can population maps as accurate as census data and existing downscaling methods be constructed solely from MP data, but these data offer additional benefits in terms of measuring population dynamics. Further, as highlighted in *SI Appendix*, section A.3, a combination of both the MP and RS methods facilitates the improvement of both spatial and temporal resolutions and demonstrates how high-resolution population datasets can be produced for any time period.

In countries where detailed human population census data are available at high resolution, the main value added is not so much in the gain in spatial resolution, but more in the ability to estimate population numbers and densities at high spatial resolution for any time period. This ability allows us to follow how population distribution changes through time in relation to the week, the season, or any particular event affecting populations over large spatial extents. The relevance of the MP approach is even greater in low-income countries where population distribution data may be scarce, outdated, and unreliable. In Africa, great variation exists in the quality of spatially referenced population data. In Malawi for example, censuses have been performed once per decade for the past three decades and data are readily available at the level of enumeration areas (i.e., administrative units of 9.38 km^2 on average). In contrast, in the Democratic

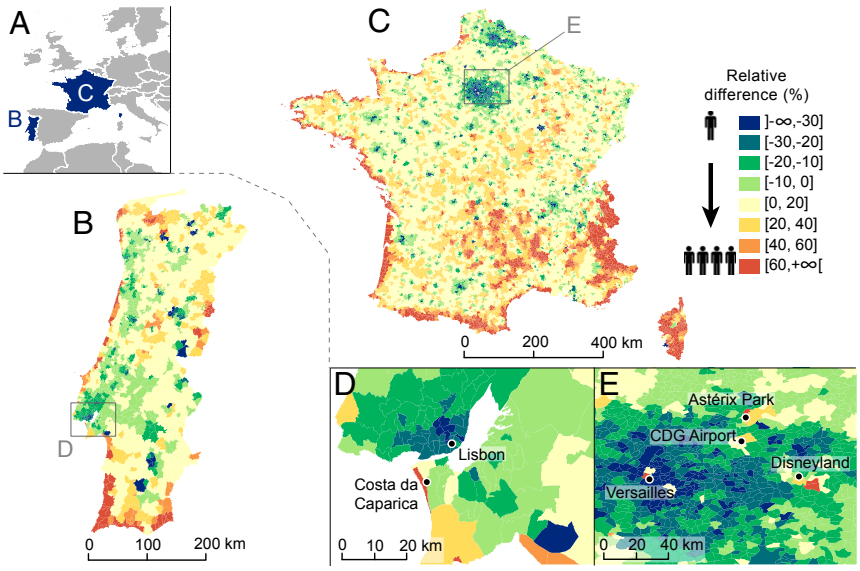

Fig. 3. Seasonal changes in population distribution in Portugal and France. (*A*) Location of Portugal and France in western Europe. (*B–E*) Relative difference in predicted population density between the main holiday period (July and August) and the working period (September to June) by administrative unit level 5 (ADM-5) in (*B*) continental Portugal and (*C*) metropolitan France. (*D*) Close-up around Lisbon with labels showing the city center of Lisbon and the seaside resort Costa da Caparica. (*E*) Close-up around Paris with labels showing the busiest airport in the country (Paris Charles de Gaulle), one of the most visited places in France (Palace of Versailles), and two popular recreation areas (Disneyland and Asterix Park).

Republic of the Congo (DRC), the most recent census was undertaken in 1984 and data are available only at the level of territories (i.e., administrative units of 12,466 km^2 on average). However, in the DRC, the MP penetration rate, although biased toward certain demographic groups, is relatively high [69% on average by the end of 2014 in Africa (31)], and the MP approach would produce considerable improvements in current knowledge of how population is distributed in the country. Even if at present the most remote and isolated populations may not have reception in some low-income countries, possibly affecting the ability to produce a comprehensive countrywide map, network coverage continues to grow at a rapid rate everywhere.

Applying the approach to countries such as the DRC, where reliable training data may not be available, requires some adjustments and assumptions, particularly regarding the relation between the MP user density and the population density, through estimates for the parameters α and β. This relation indeed may vary among and within countries according to the penetration rate of the network operator and phone use behaviors. Network access costs and cultural differences among countries may, for instance, result in communication via text messages being preferred over calls in some countries. Such differential phone use among countries might largely be accounted for by adjusting total populations by using national population counts. A further complication is that phone use and penetration rates rarely are uniform within countries. In France, the general penetration rate varies from 62.8 in the Franche-Comté region to 117.9 in Ile-de-France, according to the Autorité de Régulation des Communications Electroniques et des Postes (www.arcep.fr; accessed February 2, 2014). Such regional MP ownership information generally is available either from independent bodies such as regulators or phone operators themselves, or may be estimated through national household surveys, such as the Demographic and Health Surveys (dhsprogram.com; accessed April 1, 2014), and give a first indication of potential phone use variations among regions. The spatially stratified cross-validation procedure used here enables assessment of the impact of regional variations on model parameters in Portugal (*SI Appendix*, section B) and France (*SI Appendix*, section C.4), as well as the impact of such variation on population mapping accuracies (*SI Appendix*, section B.3). Spatial variations in phone use behaviors also may be the result of economic, social, demographic, or cultural characteristics that may be spatially clustered, therefore biasing population density estimates. Although a complete analysis of such potential biases is beyond the scope of this study, here we showed that phone use behaviors were relatively stable across space and time in Portugal and that a large part of the variation is correlated with population density and therefore is captured by the coefficient β (*SI Appendix*, section C.3).

To be applied widely and to facilitate the acquisition of MP data, the method outlined here may be simplified by using the density of phone calls instead of the density of different users over a certain time window. Even if the resulting population density datasets are marginally less accurate, this approach allows the method to become independent from user identifier data and further reduces privacy concerns (*SI Appendix*, section C.2). Similarly, using daily-aggregated data instead of night data again reduces the accuracy of estimates marginally, although notably simplifying the acquisition and processing of MP data.

The observed robustness of the MP method offers promise for extension of the mapping to other countries and network providers. However, applying the method to low-income countries where penetration rates are increasing rapidly but still exclude an important fraction of the population would require further sensitivity analyses of the impact of phone use inequalities, especially as marginalized populations also are the most vulnerable to disasters, outbreaks, and conflicts. Mobility estimates in Kenya were found to be surprisingly robust to the substantial biases in phone ownership across different geographical and socioeconomic groups (65), but these results would need to be confirmed for population density estimates.

Mobile phone call data records are collected constantly by network providers, but the potential of such data is demonstrated only sporadically. A wider use of such data currently is impeded principally by privacy and data access concerns. The use of call data records does raise important privacy concerns linked to fundamental questions of personal freedom and ethics. Studies of individual mobility patterns provide little anonymity, as the movements of individuals can be reconstructed in time and space, even if spatially and temporally coarsened datasets are used (66). Here, by using only phone call activity aggregated by towers, neither individual data nor connections between towers are used, guaranteeing the privacy of MP users. A facilitated access to anonymized and aggregated forms of these data would greatly improve our knowledge of human population distributions and movements. Network providers sometimes are reticent to share their data because of privacy and marketing concerns.

However, this study has shown that aggregated and anonymized MP data might cost-effectively provide accurate maps of population distribution for every country in the world for every month. Partnerships between governments and phone companies supported by appropriate incentives might enable fast and cheap production of population maps in emergency contexts, enabling rapid assessments of populations at risk or those affected by disasters, disease outbreaks, or conflict.

Materials and Methods

MP and Population Data. Two large datasets of MP calls obtained from major carriers in Portugal and France were used as proxies for population activity in the countries. Datasets cover the following periods: July to August 2007 and November 2007 to June 2008 (10 mo) for Portugal and May to October 2007 (5 mo) for France. Both datasets contain more than a billion calls from 2 million users in Portugal (~20% of the total population) and 17 million users in France (~30% of the total population). According to the operators, their penetration rates were uniform over the country at the time. Only calls were considered here; text messages were excluded. MP contracts from companies were removed from both datasets to include only MP contracts of individuals. For each call, the originating and receiving towers and the day the call was made were obtained. In addition, the time the call was made and a user identifier were available for Portugal only. All data used in this study can be obtained for the replication of results by contacting the corresponding author and are subject to the mobile phone carrier's nondisclosure agreement.

Census population data were obtained from the National Institute of Statistics of Portugal for 2011 (www.ine.pt; accessed January 30, 2014) and from the National Institute of Statistics and Economic Studies of France for 2007 (www.insee.fr; accessed January 30, 2014). Census population data were matched to administrative units with identifier codes. For both countries, the finest administrative unit level available (ADM-5) was used, which corresponds to "Freguesias" in Portugal (n = 2,882) and "Communes" in France (n = 36,610). The spatial resolution of administrative units is similar in France and Portugal, with average spatial resolutions (i.e., square root of the land area divided by the number of administrative units) of 3.9 km and 5.6 km, respectively.

Mapping People Based on MP Data. For each MP tower j in Portugal, we know the total number of different users T_j who made or received phone calls from/to that tower. When one makes a phone call, the network usually identifies nearby towers and connects to the closest one (67). The coverage area of a tower j thus was approximated by using a Voronoi-like tessellation (68). The Voronoi polygon associated with tower j is denoted v_j. The MP user density of the polygon v_j, denoted as σ_{v_j}, then is equal to T_j/A_{v_j}, where A_{v_j} is the area of the Voronoi polygon corresponding to tower j. An illustration of these polygons derived from MP towers is given in *SI Appendix*, section A.2.

The estimation of the population density for an administrative unit c_i based on the MP user density σ_{v_j} is a two-step method. First, the nighttime (i.e., from 8:00 PM to 7:00 AM) MP user density σ_{c_i} for c_i is computed with the following equation:

$$\sigma_{c_i} = \frac{1}{A_{c_i}} \sum_{v_j} \sigma_{v_j} A_{(c_i \cap v_j)}, \quad [1]$$

where A_{c_i} is the area of administrative unit c_i and $A_{(c_i \cap v_j)}$ is the intersection area of c_i and the Voronoi polygon v_j.

Second, nighttime MP user density values σ_{c_i} assigned to each administrative unit were compared with baseline census-derived population densities available in a training set, denoted as ρ_{c_i}. Our approach is modeled as follows:

$$\rho_c = \alpha \sigma_c^{\beta}, \quad [2]$$

where $\rho_c = [\rho_{c_1}, \rho_{c_2}, \ldots, \rho_{c_n}]$ and $\sigma_c = [\sigma_{c_1}, \sigma_{c_2}, \ldots, \sigma_{c_m}]$. The parameter α represents the scale ratio and β the superlinear effect of population density ρ_c on the nighttime MP user density σ_c. This can be transformed to $\log(\rho_c) = \log(\alpha) + \beta \log(\sigma_c)$, where a standard linear regression model with population-weighted least squares was applied to estimate the two parameters α and β. The variability of α and β was assessed using standard and spatially stratified cross-validation procedures (*SI Appendix*, section B.1). Nighttime population densities $\tilde{\rho}_c$ of all administrative units were estimated using Eq. **2**, and the total population approximation $\hat{P}$ was extracted. Nighttime population densities $\tilde{\rho}_c$ then were adjusted to make the total estimated population match the census-derived national population P:

$$\rho_c = \frac{P}{\hat{P}} \alpha \sigma_c^{\beta}. \quad [3]$$

Comparison with the RS Method. To assess the accuracy and precision of the MP method described above, we produced a nighttime population map based on a recently developed dasymetric modeling approach that incorporates a wide range of remotely sensed and geospatial data (called the RS method in this paper; *SI Appendix*, section A.1). Ancillary data layers were used, including the Corine Land Cover 2006 dataset (69), OpenStreetMap-derived infrastructure (70), satellite nightlights (71), and slope (72), among others (19). The method combines data in a flexible "Random Forest" model to generate gridded predictions of population density at ~100 m spatial resolution (*SI Appendix*, section A.1) (19). Analyses have shown that this algorithm produces improved mapping accuracies compared with previous approaches (19). The output prediction layer was used as the weighting surface to perform dasymetric redistribution of the census counts at a country level as follows (*SI Appendix*, section A.2):

$$\rho_i^{RS} = \frac{w_i}{\sum_j w_j} P, \quad [4]$$

where ρ_i^{RS} is the population density in pixel i estimated by the RS method, w_i is the weight assigned to pixel i, and P is the total population.

For comparative purposes, the same spatially stratified training dataset ("Norte" region) was used to estimate nighttime population densities in both the MP and RS methods. To assess the precision and the accuracy of the different population downscaling methods, we extracted the average nighttime population density within each of the finest level census units (ADM-5) as estimated by both methods and compared it with the baseline census-derived population densities (ρ_c) within each unit by using the Pearson product–moment correlation coefficient (r) and rmse.

Extrapolation Capacity. To further explore the stability of population density estimates derived from MP data and the capacity of extrapolation to data-scarce countries, the method was applied to the France dataset. Here, only the daily aggregated phone call activity at each tower was used, without any individual information and without the time of phone calls. This approach had two benefits: (*i*) it ensured that our population density estimation method required only data that were collected readily and stored by network providers for billing purposes and (*ii*) the privacy of network customers was preserved further. Uncertainties associated with the use of phone call densities instead of user densities and daily-aggregated MP data instead of nighttime MP data are evaluated in *SI Appendix*, section C.2.

Dynamic Mapping of Population Distributions. Temporal dynamics were derived from MP data by using the timestamp associated with each MP call. Daily dynamics were analyzed by dividing the MP data into calls performed during the day (7:00 AM to 8:00 PM) and the night (8:00 PM to 7:00 AM). Weekly dynamics were analyzed by dividing the MP data into calls performed during weekdays (Monday to Friday) and calls performed during weekends (Saturday and Sunday). Seasonal dynamics were analyzed by dividing MP data into calls performed during the holiday period (July and August) and calls performed during working periods (all other months). Predicted population densities for each unit and for both time periods were computed using best-fit α and β estimates, and relative differences between the two time periods were extracted.

ACKNOWLEDGMENTS. We thank three anonymous referees for their useful comments on an earlier version of this paper. P.D., C.L. and M.G. are supported by the Fonds National de la Recherche Scientifique (FNRS); part of this work was supported by the FNRS (PDR T.0073.13). A.J.T. is supported by funding from the NIH/National Institute of Allergy and Infectious Diseases (U19AI089674), the Bill & Melinda Gates Foundation (OPP1106427,1032350), and the Research and Policy for Infectious Disease Dynamics program of the Science and Technology Directorate, Department of Homeland Security, and Fogarty International Center, NIH. This work forms part of the WorldPop Project (www.worldpop.org.uk) and Flowminder Foundation (www.flowminder.org).

1. Tatem A, Linard C (2011) Population mapping of poor countries. *Nature* 474(7349):36–36.
2. Bongaarts J, Sinding S (2011) Population policy in transition in the developing world. *Science* 333(6042):574–576.
3. Tatem AJ, et al. (2013) Millennium development health metrics: Where do Africa's children and women of childbearing age live? *Popul Health Metr* 11(1).
4. Checchi F, Stewart BT, Palmer JJ, Grundy C (2013) Validity and feasibility of a satellite imagery-based method for rapid estimation of displaced populations. *Int J Health Geogr* 12(4).
5. Linard C, Tatem AJ (2012) Large-scale spatial population databases in infectious disease research. *Int J Health Geogr* 11(7).
6. O'Neill BC, et al. (2010) Global demographic trends and future carbon emissions. *Proc Natl Acad Sci USA* 107(41):17521–17526.
7. O'Loughlin J, et al. (2012) Climate variability and conflict risk in East Africa, 1990-2009. *Proc Natl Acad Sci USA* 109(45):18344–18349.
8. Deichmann U (1996) *A Review of Spatial Population Database Design and Modeling* (National Center for Geographic Information and Analysis, Santa Barbara, CA).
9. Jones HR (1990) *Population Geography* (Guilford Press, New York).
10. Balk D, Yetman G (2004) *The Global Distribution of Population: Evaluating the Gains in Resolution Refinement* (Center for International Earth Science Information Network, New York).
11. Tobler W, Deichmann U, Gottsegen J, Maloy K (1997) World population in a grid of spherical quadrilaterals. *Int J Popul Geogr* 3(3):203–225.
12. Dobson JE, Bright EA, Coleman PR, Durfee RC, Worley BA (2000) LandScan: A global population database for estimating populations at risk. *Photogramm Eng Remote Sensing* 66(7):849–857.
13. Balk DL, et al. (2006) Determining global population distribution: methods, applications and data. *Adv Parasitol* 62:119–156.
14. Linard C, Gilbert M, Snow RW, Noor AM, Tatem AJ (2012) Population distribution, settlement patterns and accessibility across Africa in 2010. *PLoS One* 7(2):e31743.
15. Gaughan AE, Stevens FR, Linard C, Jia P, Tatem AJ (2013) High resolution population distribution maps for Southeast Asia in 2010 and 2015. *PLoS One* 8(2):e55882.
16. Azar D, Engstrom R, Graesser J, Comenetz J (2013) Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data. *Remote Sens Environ* 130:219–232.
17. Deichmann U, Balk D, Yetman G (2001) *Transforming Population Data for Interdisciplinary Usages: From Census to Grid* (Center for International Earth Science Information Network, Washington, DC).
18. Mennis J (2003) Generating surface models of population using dasymetric mapping. *Prof Geogr* 55(1):31–42.
19. Stevens FR, Gaughan AE, Linard C, Tatem AJ Disaggregating census data for population mapping using random forests with remotely-sensed and other ancillary data. *PLoS One*, in press.
20. Bhaduri B, Bright E, Coleman P, Urban M (2007) LandScan USA: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal* 69(1):103–117.
21. Azar D, et al. (2010) Spatial refinement of census population distribution using remotely sensed estimates of impervious surfaces in Haiti. *Int J Remote Sens* 31(21):5635–5655.
22. Butler D (2011) Nuclear safety: Reactors, residents and risk. *Nature* 472(7344):400–401.
23. Mondal P, Tatem AJ (2012) Uncertainties in measuring populations potentially impacted by sea level rise and coastal flooding. *PLoS One* 7(10):e48191.
24. Wegscheider S, et al. (2011) Generating tsunami risk knowledge at community level as a base for planning and implementation of risk reduction strategies. *Nat Hazards Earth Syst Sci* 11(2):249–258.
25. Jankowska MM, Lopez-Carr D, Funk C, Husak GJ, Chafe ZA (2012) Climate change and human health: Spatial modeling of water availability, malnutrition, and livelihoods in Mali, Africa. *Appl Geogr* 33:4–15.
26. Tatem AJ, Campiz N, Gething PW, Snow RW, Linard C (2011) The effects of spatial population dataset choice on estimates of population at risk of disease. *Popul Health Metr* 9(4).
27. Pindolia DK, et al. (2013) The demographics of human and malaria movement and migration patterns in East Africa. *Malar J* 12(397).
28. Tatem AJ, et al. (2012) Mapping populations at risk: Improving spatial demographic data for infectious disease modeling and metric derivation. *Popul Health Metr* 10(8).
29. Tatem AJ (2014) Mapping the denominator: Spatial demography in the measurement of progress. *Int Health* 6(3).
30. Leung S, Martin D, Cockings S (2010) Linking UK public geospatial data to build 24/7 space-time specific population surface models. *GIScience 2010: Sixth International Conference on Geographic Information Science* (Springer LNCS, Zurich).
31. International Telecommunication Union (2014) World Telecommunication Development Conference (WTDC-2014): Final Report. (ITU, Dubai, United Arab Emirates).
32. Mateos P, Fisher PF (2006) Spatiotemporal accuracy in mobile phone location: Assessing the new cellular geography. *Dynamic and Mobile GIS: Investigating Change in Space and Time*, eds Drummond J, Billen R, João E, Forrest D (Taylor & Francis, Boca Raton, FL), pp 189–212.
33. Watts DJ (2007) A twenty-first century science. *Nature* 445(7127):489.
34. Lazer D, et al. (2009) Social science. Computational social science. *Science* 323(5915):721–723.
35. Vespignani A (2009) Predicting the behavior of techno-social systems. *Science* 325(5939):425–428.
36. Blondel VD, et al. (2013) Data for development: the D4D challenge on mobile phone data. arXiv:12100137v2.
37. González MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. *Nature* 453(7196):779–782.
38. Song C, Qu Z, Blumm N, Barabási A-L (2010) Limits of predictability in human mobility. *Science* 327(5968):1018–1021.
39. Lu X, Wetter E, Bharti N, Tatem AJ, Bengtsson L (2013) Approaching the limit of predictability in human mobility. *Sci Rep* 3:2923.
40. Järv O, Ahas R, Saluveer E, Derudder B, Witlox F (2012) Mobile phones in a traffic flow: A geographical perspective to evening rush hour traffic analysis using call detail records. *PLoS One* 7(11):e49171.
41. Ratti C, Pulselli RM, Williams S, Frenchman D (2006) Mobile landscapes: Using location data from cell phones for urban analysis. *Environ Plann B Plann Des* 33(5):727–748.
42. Pulselli R, Ramono P, Ratti C, Tiezzi E (2008) Computing urban mobile landscapes through monitoring population density based on cellphone chatting. *Int J Des Nat Ecodynamics* 3(2):121–134.
43. Reades J, Calabrese F, Ratti C (2009) Eigenplaces: Analysing cities using the space–time structure of the mobile phone network. *Environ Plann B Plann Des* 36(5):824–836.
44. Lenormand M, et al. (2014) Cross-checking different sources of mobility information. *PLoS One* 9(8):e105184.
45. Ahas R, Aasa A, Mark Ü, Pae T, Kull A (2007) Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tour Manage* 28(3):898–910.
46. Ahas R, Aasa A, Roose A, Mark Ü, Silm S (2008) Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tour Manage* 29(3):469–486.
47. Bengtsson L, Lu X, Thorson A, Garfield R, von Schreeb J (2011) Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in Haiti. *PLoS Med* 8(8):e1001083.
48. Lu X, Bengtsson L, Holme P (2012) Predictability of population displacement after the 2010 Haiti earthquake. *Proc Natl Acad Sci USA* 109(29):11576–11581.
49. Wesolowski A, et al. (2012) Quantifying the impact of human mobility on malaria. *Science* 338(6104):267–270.
50. Tatem AJ, et al. (2009) The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents. *Malar J* 8:287.
51. Tatem AJ, et al. (2014) Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning. *Malar J* 13:52.
52. Schläpfer M, et al. (2014) The scaling of human interactions with city size. *J R Soc Interface* 11(98):20130789.
53. Gomez-Lievano A, Youn H, Bettencourt LMA (2012) The statistics of urban scaling and their connection to Zipf's law. *PLoS One* 7(7):e40393.
54. Krings G, Karsai M, Bernhardsson S, Blondel VD, Saramäki J (2012) Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Sci* 1(1):1–16.
55. Mitzenmacher M (2004) A brief history of generative models for power law and lognormal distributions. *Internet Math* 1(2):226–251.
56. Bahn V, McGill BJ (2013) Testing the predictive performance of distribution models. *Oikos* 122(3):321–331.
57. Wenger SJ, Olden JD (2012) Assessing transferability of ecological models: An underappreciated aspect of statistical validation. *Methods Ecol Evol* 3(2):260–267.
58. Gilbert M, et al. (2014) Predicting the risk of avian influenza A H7N9 infection in live-poultry markets across Asia. *Nat Commun* 5(4116).
59. Graham M, Shelton T (2013) Geography and the future of big data, big data and the future of geography. *Dialogues Hum Geogr* 3(3):255–261.
60. Goodchild MF (2013) The quality of big (geo)data. *Dialogues Hum Geogr* 3(3):280–284.
61. Eubank S, et al. (2004) Modelling disease outbreaks in realistic urban social networks. *Nature* 429(6988):180–184.
62. Colizza V, Barrat A, Barthélemy M, Vespignani A (2006) The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc Natl Acad Sci USA* 103(7):2015–2020.
63. Bohorquez JC, Gourley S, Dixon AR, Spagat M, Johnson NF (2009) Common ecology quantifies human insurgency. *Nature* 462(7275):911–914.
64. Bagrow JP, Wang D, Barabási A-L (2011) Collective response of human populations to large-scale emergencies. *PLoS One* 6(3):e17680.
65. Wesolowski A, Eagle N, Noor AM, Snow RW, Buckee CO (2013) The impact of biases in mobile phone ownership on estimates of human mobility. *J R Soc Interface* 10(81):20120986.
66. De Montjoye Y-A, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the crowd: The privacy bounds of human mobility. *Sci Rep* 3(1376).
67. 3GPP (2000) TS 03.02 V7.1.0 network architecture. Available at www.3gpp.org/ftp/Specs/html-info/0302.htm. Accessed February 4, 2014.
68. Okabe A, Boots B, Sugihara K, Chiu SN (2000) *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams* (Wiley, New York).
69. European Environment Agency (2013) Corine Land Cover 2006 raster data, version 17. Available at www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-3. Accessed September 16, 2013.
70. OpenStreetMap Contributors (2013) OpenStreetMap database. Available at OpenStreetMap.org. Accessed May 19, 2014.
71. National Oceanic and Atmospheric Association, National Geophysical Data Center (2012) Earth Observation Group: VIIRS nighttime lights—2012. Available at www.ngdc.noaa.gov/dmsp/data/viirs_fire/viirs_html/viirs_ntl.html. Accessed August 4, 2013.
72. Lehner B, Verdin K, Jarvis A, Fund WW (2006) HydroSHEDS Technical Documentation. World Wildlife Fund US, Washington, DC. Available at hydrosheds.cr.usgs.gov. Accessed May 22, 2014.

Supplementary material

Table of Contents

A. Comparison of methods for mapping human population density

A.1. Mapping human population density using remotely-sensed and other geospatial data (RS method) 2

A.2. Schematic illustrations of population density estimation methods 3

A.3. Combination of MP and RS methods 3

B. Variability of parameters α and β

B.1. Cross-validation procedures 6

B.2. Variability of α and β according to the cross-validation procedure 7

B.3. Sensitivity analysis of population estimates to α and β 8

C. Flexibility, potential bias and extrapolation capacity

C. 1. Density of MP towers 10

C.2. Daily aggregated data and density of MP calls 11

C.3. Spatio-temporal variability in phone usage 14

C.4. Application to France 18

D. Population dynamics 20

References 23

A. Comparison of methods for mapping human population density

In addition to the MP-based mapping, human population densities were predicted using more traditional modelling methods developed by the WorldPop project[1]. A semi-automated dasymetric modelling approach that incorporates census and ancillary data layers in a flexible Random Forest statistical model was applied to generate gridded predictions of population density at approximately 100m spatial resolution (1). The combination of satellite and other geospatial datasets in a Random Forest framework has been shown to produce substantial increases in population mapping accuracies over previous approaches (1).

A.1. Mapping human population density using remotely-sensed and other geospatial data (RS method)

Ancillary data layers used as covariates include the CORINE Land Cover 2006 dataset[2], OpenStreetMap-derived infrastructure[3], satellite nightlights[4], slope[5], amongst others related to human population distributions. All data were processed to ensure that projections, resolutions, and extents matched. The method combines data in a Random Forest model to generate gridded predictions of population density at ≈100 m spatial resolution ($8.33*10^{-4}$ decimal degrees). The Random Forest model is an ensemble, nonparametric approach that generates multiple individual classification or regression trees, and from which a final prediction is made based on an average of the prediction estimates from individual regression trees (2, 3). By using an ensemble of trees, the Random Forest approach provides flexibility for both continuous and discrete data and both linear and non-linear relationships between predictor and response variables. These predictors may be included in different combinations across the many regression trees in the forest, chosen at random and used to estimate an output weighting layer using only the combinations proven to increase out-of-bag prediction accuracy. The model is parameterized by aggregating covariates by administrative units (from the training dataset) and using them in a semi-automated Random Forest predictive model (2, 3) to estimate a population density weighting layer at a spatial resolution of 100 m. This prediction layer was then used as the weighting surface to perform a dasymetric redistribution of the national population to create a population density surface. Model estimation, fitting and prediction were completed using the statistical environment R 3.0.1 (4) and the randomForest package 4.6-7 (3).

[1] WorlPop project: www.worldpop.org.uk [Accessed April 1, 2014]
[2] European Environment Agency (2013) Corine Land Cover 2006 raster data, version 17. Available at: http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-3 [Accessed September 16, 2013]
[3] http://www.openstreetmap.org/ [Accessed September 12, 2013]
[4] http://ngdc.noaa.gov/eog/viirs/download_viirs_ntl.html [Accessed January 20, 2014]
[5] http://hydrosheds.cr.usgs.gov/index.php [Accessed January 20, 2014]

A.2. Schematic illustrations of population density estimation methods

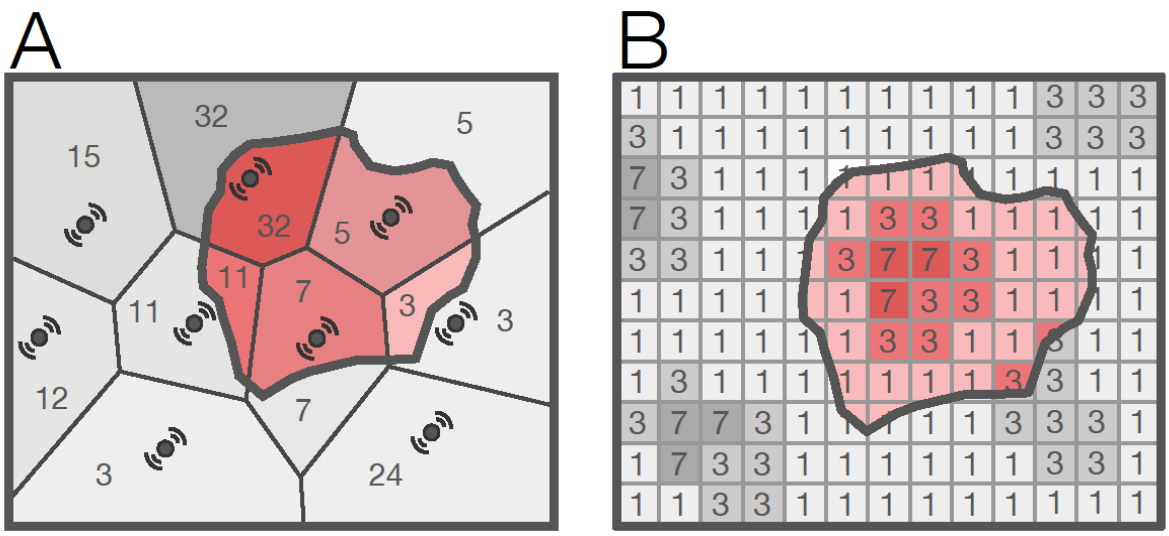

Figure S1: (A) Illustration of the *MP* method, where Voronoi polygons are built based on the spatial configuration of MP towers. The MP call density of an area (red polygon) is derived from the proportion of Voronoi polygons intersecting that area, as described in Equation 1, and the population density is a function of the MP user density at night (Equation 2). (B) Illustration of the *RS* method, where a relative weight is assigned to each pixel according to its environmental and infrastructural characteristics. The estimated population density of a commune (red polygon) is given by the average population density of pixels that fall within the commune.

A.3. Combination of MP and RS methods

In order to optimize both spatial and temporal resolutions, the *MP* method developed in the main paper can be combined with the *RS* approach described above. In a first step, we estimated the nighttime population of each Voronoi polygon $\boldsymbol{v_j}$ that corresponds to the coverage area of tower *j*. Then, the population of $\boldsymbol{v_j}$ is disaggregated to ≈100m grid squares using the Random Forest approach described in section A.1. The combination of both methods (*COMB*) captures the spatial details resulting from the *RS* method, especially in more rural areas where the density of MP towers is low, and the spatial details resulting from the *MP* method, especially in urban areas where the distance between MP towers is often finer than the spatial resolution of the geospatial datasets used in the *RS* method (Fig. S2). Here we used the same training (*Norte* region) and evaluation datasets as in Figure 2 of the main manuscript and extracted accuracy statistics. An overall higher accuracy is achieved with the *COMB* method compared to the *MP* and *RS* methods ($RMSE^{MP}$ = 796; $RMSE^{RS}$ = 850 and $RMSE^{COMB}$ = 684), while the overall precision is identical to the *MP* method but lower than the *RS* method (r^{MP} = 0.89, r^{RS} = 0.92 and r^{COMB} = 0.89). Even though the RMSE is lower for the *COMB* method than the *RS* and *MP* methods in densely populated areas, which probably has a high impact on the global RMSE, Fig. S3 shows that the *COMB* method produced less accurate results for a large part of the lower population density classes. This is mainly due to discrepancies between the distribution of MP user at night and census-derived (i.e. residential) population distribution due for example to a higher density of MP users along roads.

Note that a few minor improvements such as prohibiting population from water and other uninhabited regions are straightforward and would marginally increase the accuracy of the MP method.

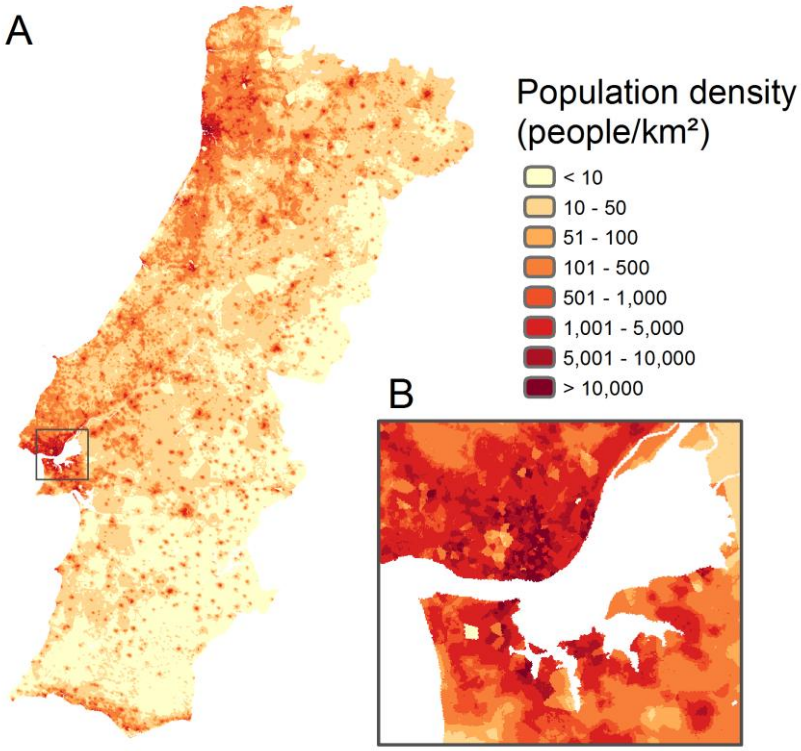

Figure S2: Population density at 100 x 100 m spatial resolution, as estimated by the combination of the *MP* and *RS* methods: (A) mainland Portugal with (B) close-up around the capital city Lisbon.

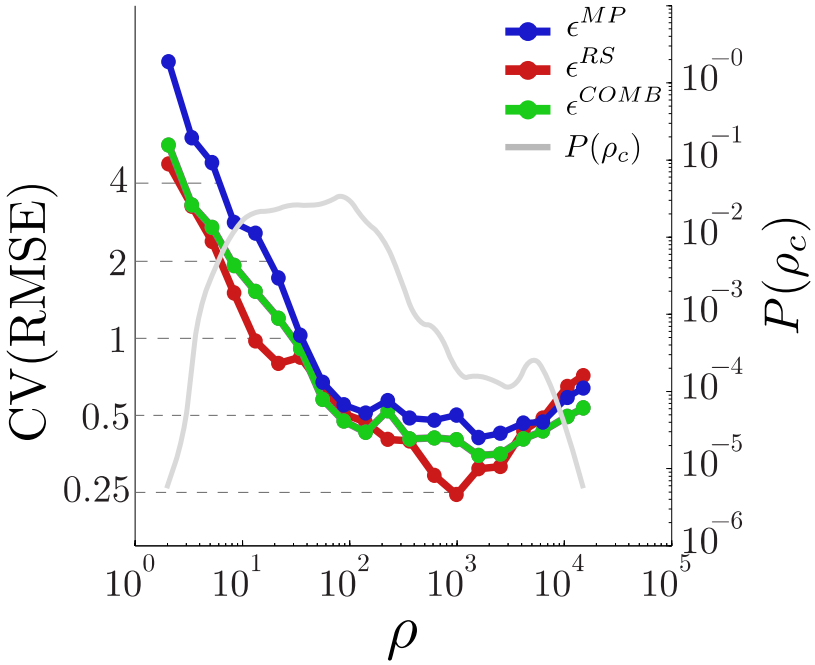

Figure S3: RMSEs normalized by the average population density of intervals, for the *MP* (blue), *RS* (red) and *COMB* methods (green). To aid visualisation, RMSEs are plotted on a logarithmic scale. The grey line represents the absolute population summed by population density intervals.

B. Variability of parameters α and β

Understanding and quantifying the stability of the estimated parameters α and β is important for the method presented in the main paper to be applied elsewhere. As outlined in Equation 2, α and β were estimated by using a linear regression on training data to model the relation between MP call density and population density in each commune. Choosing one particular training set over another can lead to different estimations of the parameters as different human behaviours or penetration rates can be observed across regions (5).

Two types of cross-validation procedures are presented here: a standard and a spatially-stratified cross-validation procedure (section B.1.). The range of values obtained for α and β (section B.2.) was then used to test the sensitivity of population density estimations to these parameters (section B.3.).

B.1. Cross-validation procedures

In the standard cross-validation procedure, 30% of administrative units were randomly sampled and used as a training set to derive α and β coefficients. Accuracy assessment statistics (correlation *r* and RMSE) were calculated on the independent evaluation set consisting of the remaining 70% of administrative units. The sampling was repeated 1,000 times in order to provide an assessment of the variability of parameters and accuracy statistics.

Because training and evaluation records are selected at random from the dataset, and population densities are spatially correlated, even a model with poor extrapolation ability may appear to predict well when measured in this way. The ability of a model to make accurate extrapolated predictions in new locations would be better measured by performing a spatially-stratified cross-validation where training and test sets are sampled from geographically distinct regions (6).

We carried out a spatially-stratified cross-validation procedure by assigning administrative units to either the training or evaluation datasets according to whether they fell inside (training) or outside (evaluation) a disc of radius 100 km. Discs were placed at random, centred on the location of an administrative unit, subject to the constraint that the training and evaluation sets contain at least 865 administrative units (30% of the total number of administrative units in Portugal). Below this threshold, the disc radius was iteratively increased or decreased by steps of 10 km until the minimum was reached. This constraint ensured that sufficient data were available to adequately train the model and to evaluate its predictive capacity. The disc-fold validation procedure was implemented in R (4) using code adapted from the sperrorest package (7). This disc-fold validation procedure was repeated 1,000 times for each model run, and accuracy assessments were computed (correlation *r* and RMSE).

B.2. Variability of α and β according to the cross-validation procedure

The best-fit estimate of 62.95 ±2.48 was found for the parameter α when using a random cross-validation procedure, while this estimate became 69.11 ±10.49 when using a spatially-stratified cross-validation procedure (Fig. S4A). The parameter β, which captures the super linear effect that may exist between population density and MP call density, was estimated to 0.803 ±0.015 when using a standard cross-validation procedure and 0.767 ±0.055 when using a spatially-stratified cross-validation procedure (Fig. S4B). Several authors have shown that this parameter is usually slightly below 1.0 (8–11). Even though these calculations in the literature have been done on the number of calls per MP tower, and not on the density of calls in a tower's coverage area, we expected similar values in our analysis.

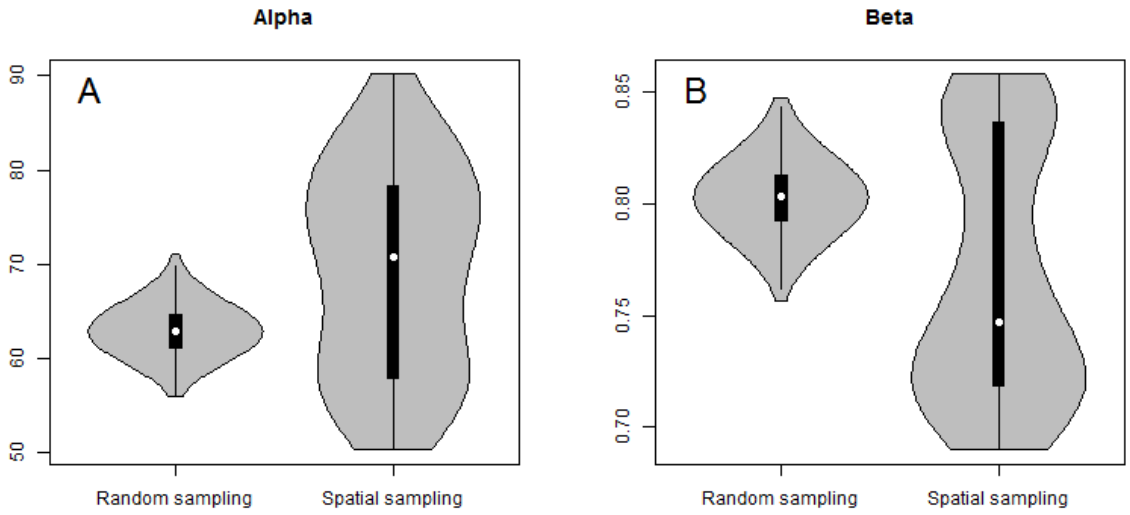

Figure S4: (A) Alpha and (B) beta coefficients estimated using randomly sampled and spatially-stratified training datasets.

While the random sampling used in the standard cross-validation procedure has the advantages of removing any cultural or economic bias existing between different geographical regions and limiting spatial autocorrelation problems in the data, the spatially-stratified cross-validation procedure enables reproduction of the initial conditions typically faced by a population distribution modeller when applying a model to a data-scarce country where detailed population data are only available for one region and the model therefore needs to be extrapolated to a geographically different region. In terms of accuracy of population density estimations, our analysis showed that the choice of a particular geographical region over another as training data may induce larger variations in global RMSE (686 ±173) than the use of a random sample of data for training (574 ±42) (Fig. S5B). Differences in correlation coefficient variations between standard and spatially-stratified cross-validation procedures are less significant, with values of 0.873 ±0.011 and 0.885 ±0.011, respectively (Fig. S5A).

When detailed training data exist for calibration, errors can be reduced by choosing a training dataset (i) representative of the larger area to be mapped and (ii) representing a large diversity of population densities. In addition, when allowed by the data, calculating different β coefficients for different regions or different population subgroups should be considered.

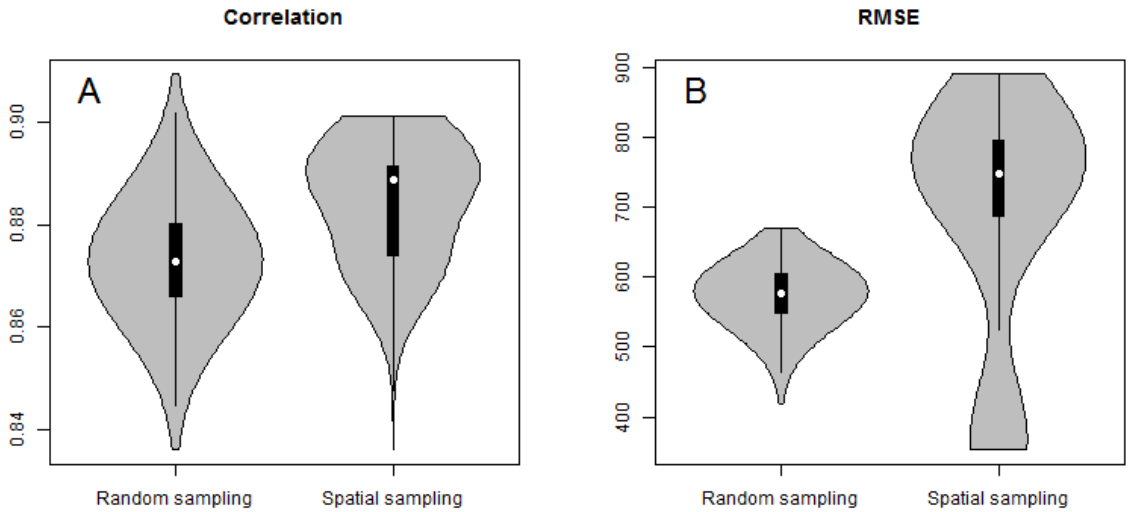

Figure S5: (A) Correlation coefficients and (B) RMSEs calculated using randomly sampled and spatially-stratified evaluation datasets.

B.3. Sensitivity analysis of population estimates to α and β

Now that we have a better idea on how α and β values may vary according to the training dataset used (see Section B.2), it is important to test the sensitivity of population density estimations to these parameters. While the variability of α might seem important, its impact on population density estimations is null, since this parameter is corrected automatically to match the total population of the country (Equation 3 in main paper). This is confirmed in Fig. S6A and S6C: when artificially changing the value of α (within the maximum range identified in previous section: 50-90), both the RMSE and the correlation coefficient *r* remain constant.

Unlike α, the sensitivity analysis shows a clear influence of β on the RMSE and *r* (Fig. S6B and S6D). A low value of the parameter β means that a proportionally lower population density is assigned to low-density areas compared to high-density areas, which can create large discrepancies in population density estimations, with overestimated population densities in urban areas and underestimated population densities in rural areas. A large value of β results in the opposite effect: overestimation of low-populated areas and underestimation of densely populated areas, resulting in an increasing global RMSE. In Figs. S6B and S6D, β values range between 0.69 and 0.86 (maximum range identified in previous section). When using values of β within the confidence interval of 0.77 ±0.055 obtained with the spatially-stratified cross-validation procedure described above, RMSE values range between 565 and 655 (15% increase) and *r* ranges between 0.88 and 0.854.

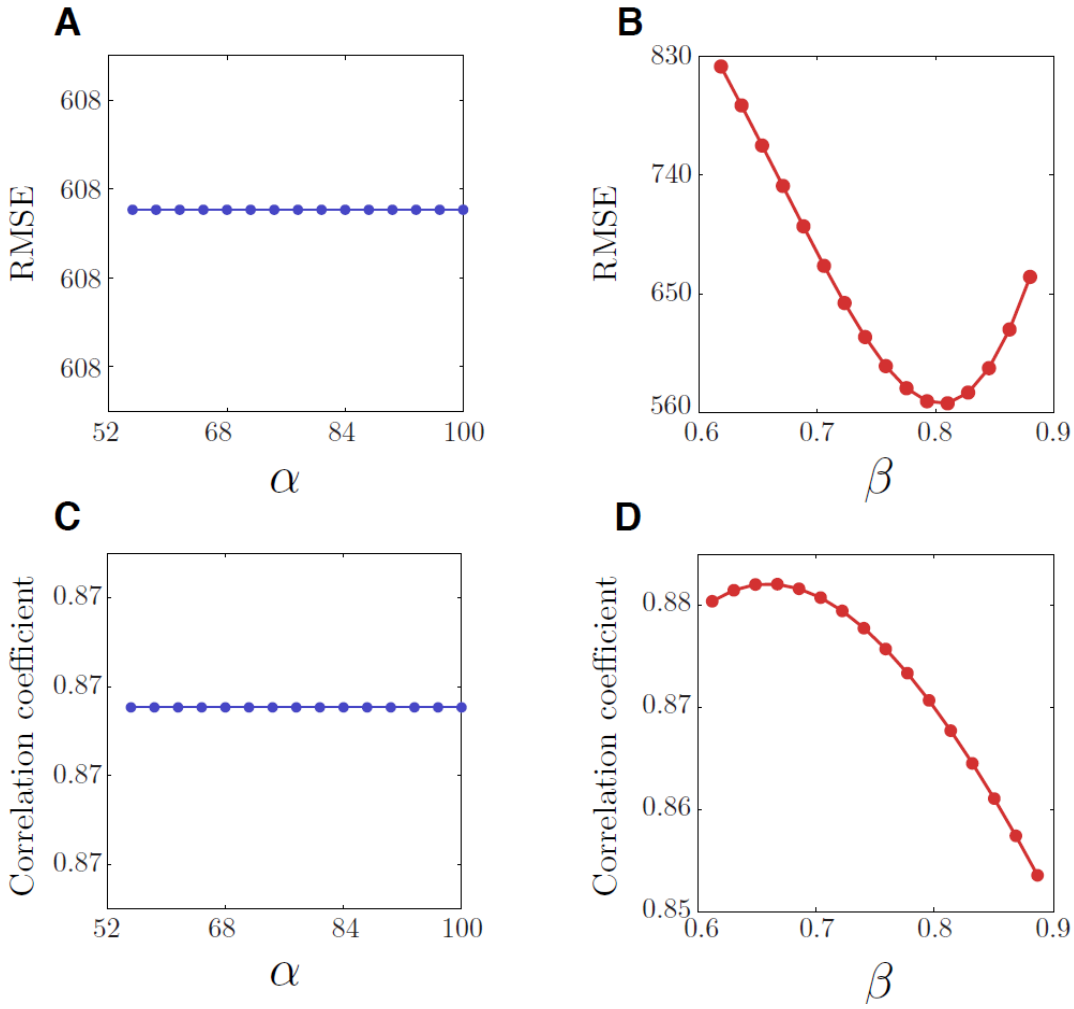

Figure S6: Influence of α and β parameters on the global RMSE and correlation coefficients.

C. Flexibility, potential bias and extrapolation capacity

In this section, we present analyses that have been done to test the flexibility of the *MP* method in terms of input data used, the impact of potential socio-economic bias and the extrapolation capacity of the method to other countries. First, we test the ability of the density of phone towers (section C.1.), the density of daily-aggregated data and the density of MP calls (section C.2.) to accurately estimate population densities. These data can often be more easily acquired from network providers than the number of MP users connected to a tower over a certain time window. The objective here is therefore to estimate the impact the use of such data would have on population estimation accuracies.

C.1. Density of MP towers

The density of MP towers by administrative unit t_{c_i} was computed with the following equation:

$$t_{c_i} = \frac{1}{A_{c_i}} \sum_{v_j} t_{v_j} A_{(c_i \cap v_j)}$$

where A_{c_i} is the area of administrative unit c_i and $A_{(c_i \cap v_j)}$ is the intersection area of commune c_i and the Voronoi polygon v_j.

In Portugal, the density of MP towers is highly correlated to census-derived population densities (r = 0.794; p < 0.0001), which suggests that using only the density of MP towers would already provide a good population density approximation (Fig. S7).

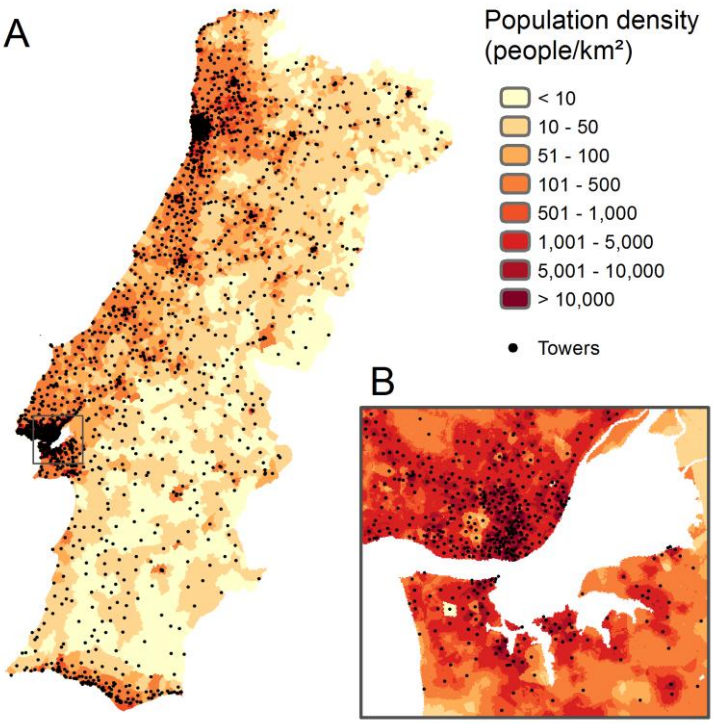

Figure S7: Spatial distribution of MP towers (A) in Portugal, with (B) close-up around the capital city Lisbon. Census-derived population densities are shown in background.

Here we compared population mapping accuracies when using the same *MP* method as described in the main paper, but using the density of MP towers instead of the density of nighttime MP users as input data. Results show that population density estimations are significantly less accurate when only using the density of MP towers (Fig. S8), with maximum RMSE values being particularly high (> 3,100) when using a spatially-stratified cross-validation procedure. In addition, the use of MP towers alone does not allow any dynamic mapping.

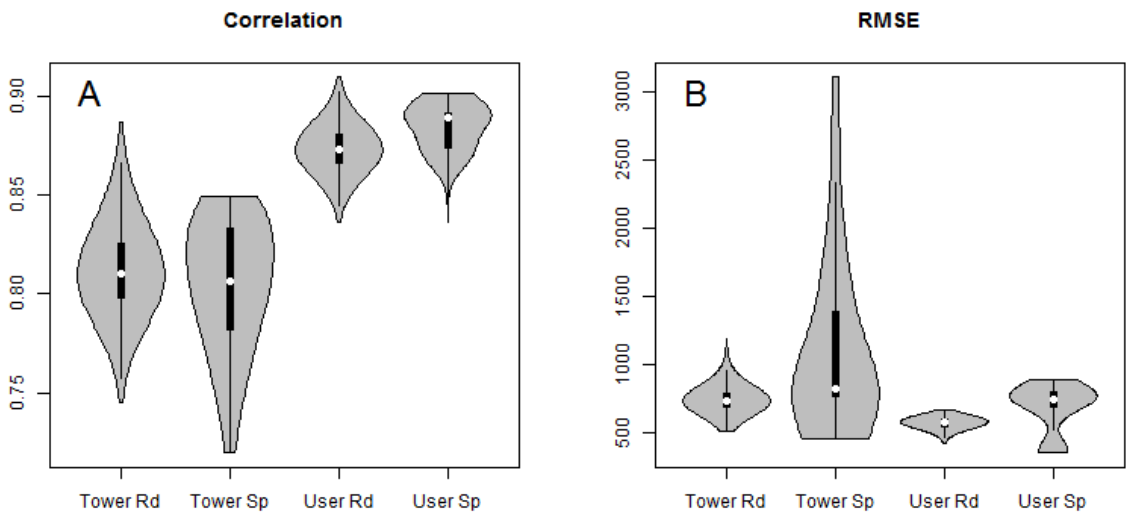

Figure S8: (A) Correlation coefficients and (B) RMSEs calculated using the density of phone towers and the density of users (Rd = standard cross-validation procedure; Sp = spatially-stratified cross-validation procedure)

C.2. Daily aggregated data and density of MP calls

The method presented in the main paper uses the density of different MP users during the night (8 p.m. - 7 a.m.) as input data. However, network providers do not always provide users' identifiers and the time of phone calls and such detailed data also reduce the level of anonymity. We therefore compared (i) the accuracy of population density datasets created from daily-aggregated data compared to nighttime data and (ii) the accuracy of datasets created from MP call data compared to MP user data. The goal is to evaluate the ability of very basic and fully-anonymized MP datasets to predict human population densities (Figs. S9 and S10).

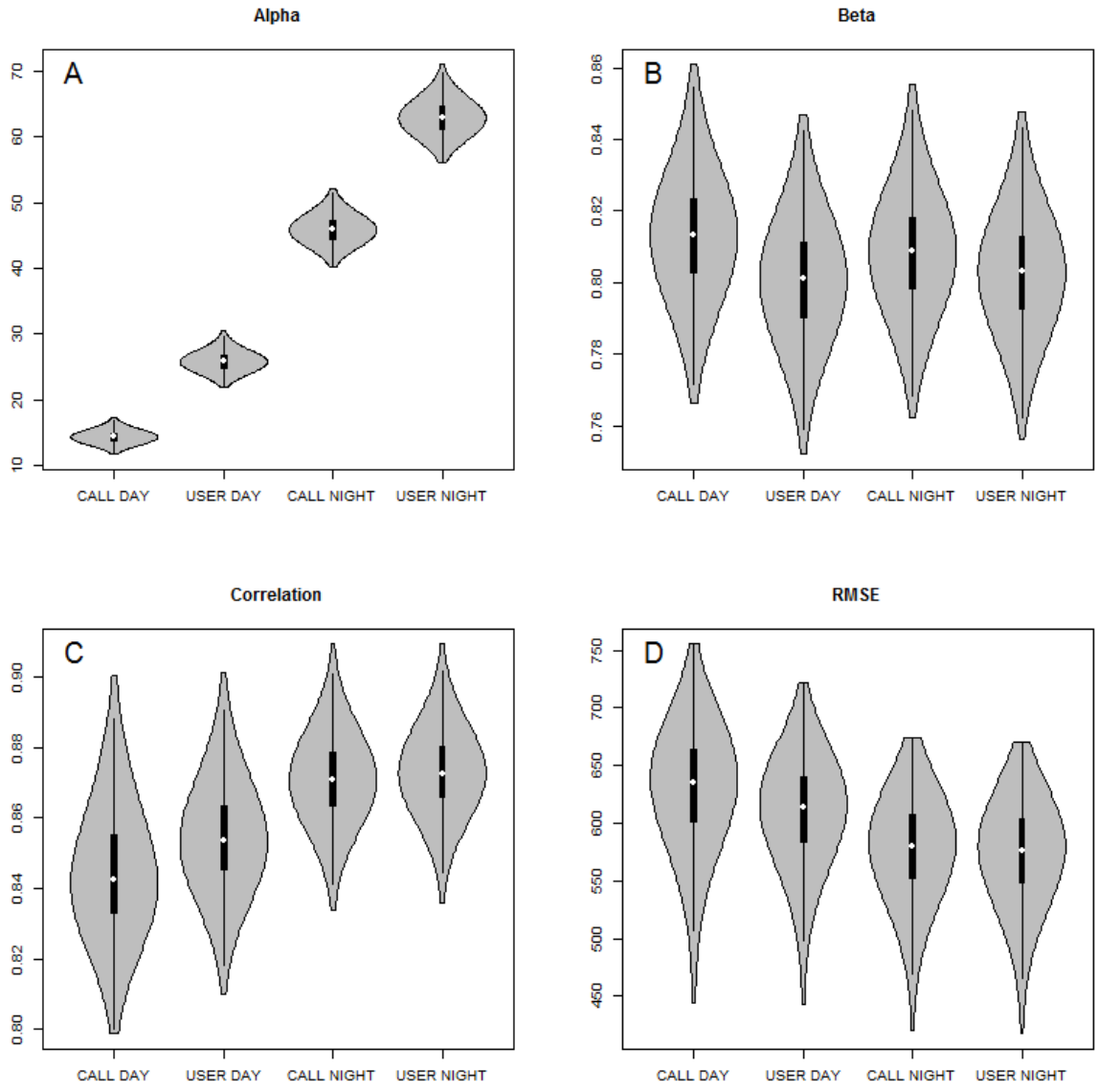

Figure S9: (A) Alpha, (B) beta, (C) correlation coefficient and (D) RMSE calculated when using (i) daily-aggregated calls (CALL DAY), (ii) daily-aggregated users (USER DAY), (iii) nighttime calls (CALL NIGHT) and (iv) nighttime users (USER NIGHT), with a standard cross-validation procedure.

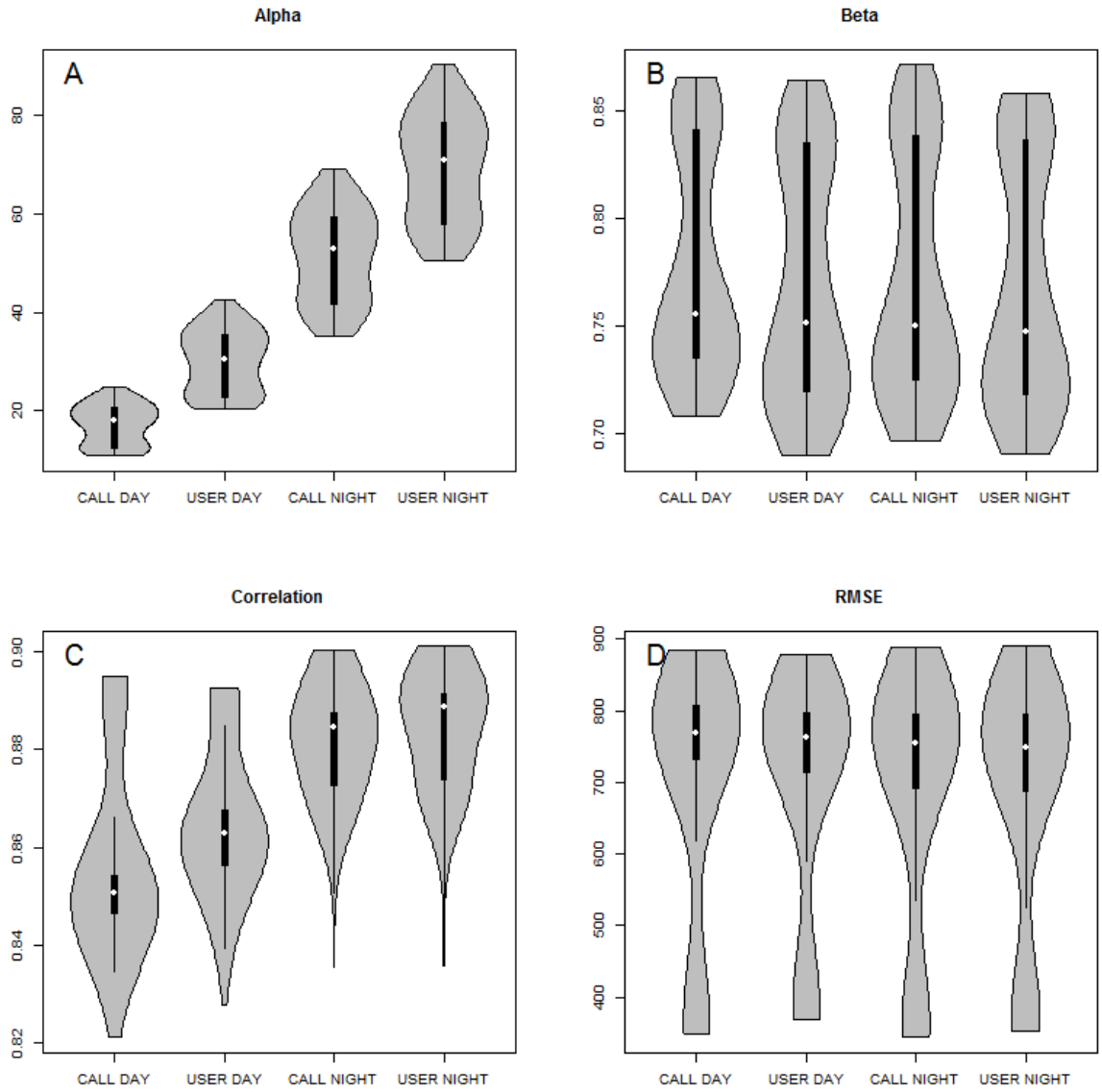

Figure S10: (A) Alpha, (B) beta, (C) correlation coefficient and (D) RMSE calculated when using (i) daily-aggregated calls (CALL DAY), (ii) daily-aggregated users (USER DAY), (iii) nighttime calls (CALL NIGHT) and (iv) nighttime users (USER NIGHT), with a spatially-stratified cross-validation procedure.

Statistical analyses including analyses of variance and Tukey's honest significant difference tests were performed to test for differences between the different datasets used as input data. The Tukey's honest significant difference statistical test is used to identify which means are significantly different from the others. This test is based on the range of the sample means rather than the individual differences.

Even if the density of calls and the density of users are very highly correlated in Portugal (r = 0.99, p < 0001), results show that population density datasets produced using the density of users are generally more precise and accurate than datasets produced using the density of calls. However,

non-significant differences in RMSE were observed between nighttime calls (CALL NIGHT) and nighttime users (USER NIGHT) when using both the standard cross-validation procedure (F=3.745; p=0.053) and the spatially-stratified cross-validation procedure (F=0.007; p=0.935), suggesting that, during the night, using the density of calls instead of the density of users does not impact significantly the accuracy of population density estimates and that the number of calls per user is relatively stable during the night.

Results also show that population density estimates produced using nighttime data were significantly more precise and more accurate than estimates produced using daily-aggregated data, with *r* and RMSE statistics being significantly different (Figs. S9 and S10). However, the accuracy assessment was done here using census-derived nighttime data as reference, which is not entirely appropriate. For a more precise accuracy assessment, we would need daytime data as reference. Nevertheless, estimated β values between both day/night and call/user data are very close (and even non-significantly different when using the spatially-stratified cross-validation procedure), which suggests a minimal impact on predicted population densities. When available MP data only include the daily-aggregated number of phone calls (without information on the number of users or on the calling time), as is the case in France, the daily-aggregated number of phone calls can reasonably replace the number of users per night, as long as phone usage behaviors are relatively stable across space and time. The spatio-temporal variability in phone usage is assessed below for Portugal.

C.3. Spatio-temporal variability in phone usage

In order to assess the variability of phone usage behaviors in time and space, MP users were divided into three distinct profiles, each containing about a third of the total number of users (Fig. S11). The profiles are based on the number of phone calls they performed at night during the studied period of 242 days: (i) Type 1 corresponding to low-activity users with less than 13 calls (0.054 per night), (ii) Type 2 corresponding to medium-activity users with number of calls between 13 and 68 ([0.054,0.28] per night), (iii) Type 3 corresponding to high-activity users with more than 68 calls (0.28 per night).

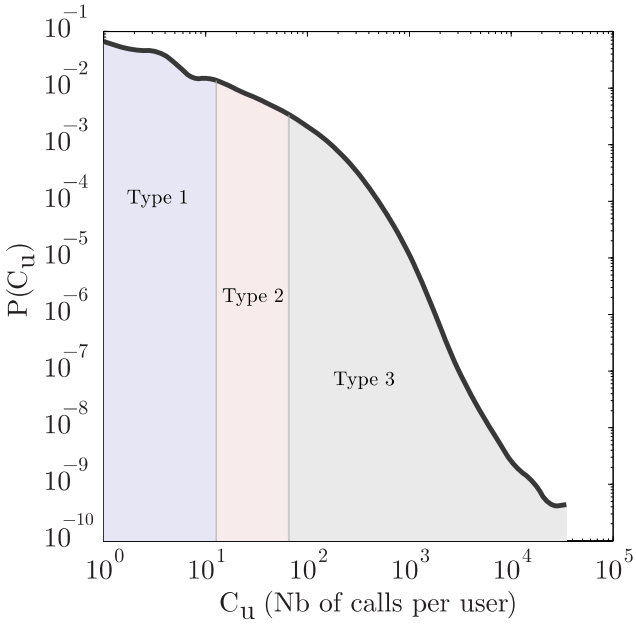

Figure S11: Probability Density Function of total number of night phone calls per user. Mobile phone users are divided into three distinct profiles, each containing a third of the users: low-activity users (Type 1), medium-activity users (Type 2) and high-activity users (Type 3).

We then analysed the variability in the proportion of users of Type 1, Type 2 and Type 3 in both time (Figs. S12 and S13) and space (Figs. S14, S15 and S16).

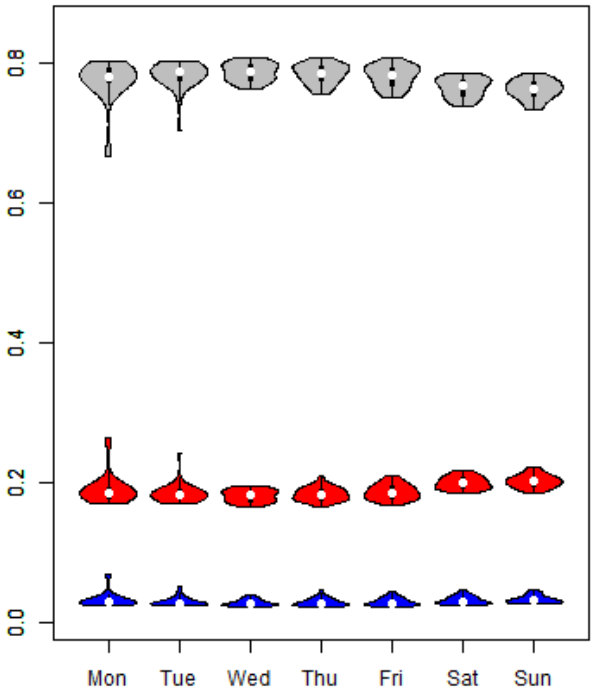

Figure S12: Variability of user profiles over time. Distribution of proportion of user of type 1 (blue), type 2 (red) and type 3 (grey) for each day of the week.

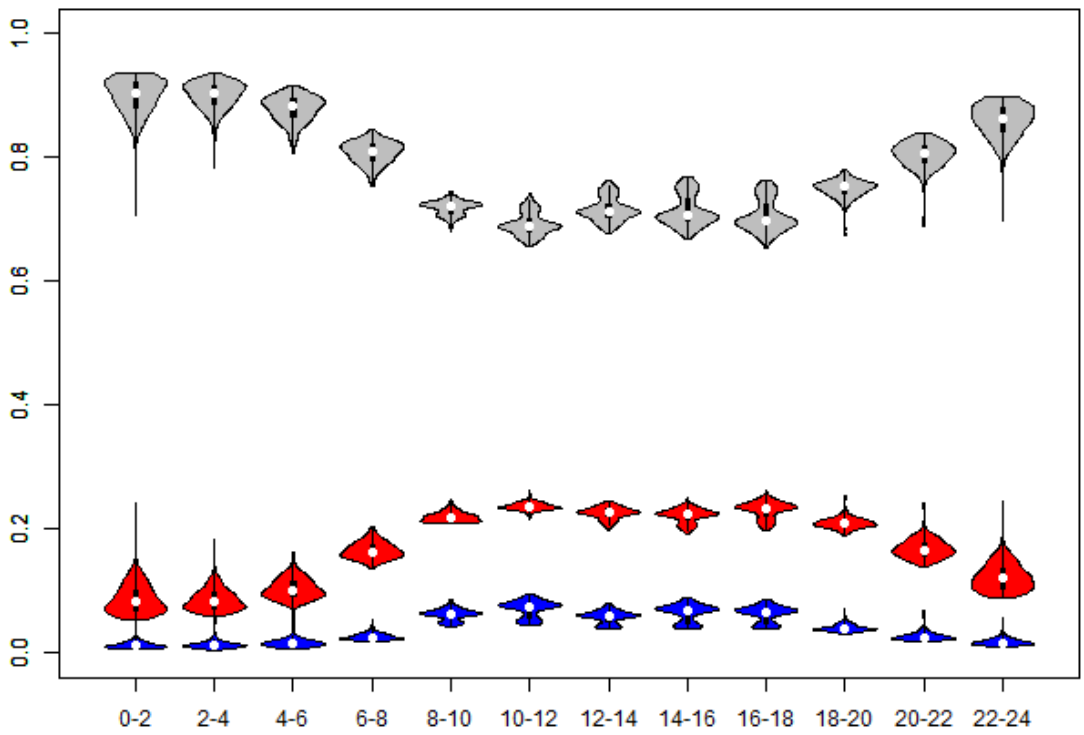

Figure S13: Variability of user profiles over time. Distribution of proportion of user of type 1 (blue), type 2 (red) and type 3 (grey) for each 2-hour period of the day.

Results show that the proportion of each profile is stable over the week (Fig. S12), but less over the day (Fig. S13). Indeed, we observe that the proportion of high-activity users (Type 3) is lower during the day than during the night while the proportion of low and medium-activity users (Types 1 and 2) is higher during the day than the night. Considering day-time and night-time data separately, as we do in our manuscript, is thus important in order to study users with stable behaviors.

To analyze the variability in the proportion of users of Type 1, Type 2 and Type 3 in space, we used three variables that are spatially clustered: the population density (Fig. S14), the unemployement rate (Fig. S15) and the percentage of people who hold a higher education degree (Fig. S16). These data were obtained from the National Institute of Statistics of Portugal by administrative unit level 5 (ADM-5) for the year 2011 (12) and were summarized by Voronoi polygon.

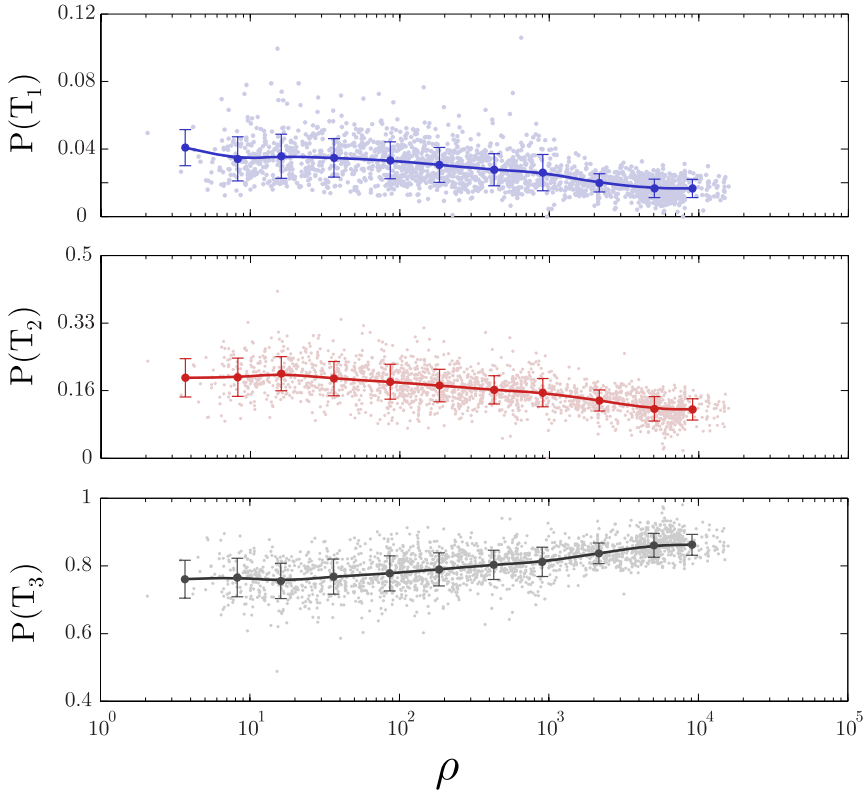

Figure S14: Variability of user profiles at each mobile phone tower over population density. The proportion of low (blue) and medium (red) activity users ($T_1$and T_2) tend to decrease in densely populated areas, while the proportion of high-activity users (grey) increases (T_3).

Fig. S14 shows that the proportion of each user profile varies across space, with a higher proportion of high activity users (Type 3) than low and medium activity users (Type 1 and 2) in densely populated areas. This well-known super-linear effect of population density on human activities is captured by the coefficient β in our model.

The proportion of each user profile also varies with the proportion of people holding a higher education degree (Fig. S15), with a larger proportion of high activity users (Type 3) in administrative units where the proportion of people holding a higher education degree is higher. However, this trend is mainly due to the correlation between the population density and the higher education degree ($r = 0.52$; $p < 0.0001$), which suggests that the influence of the education level is captured by the coefficient β. There is however no clear relation in the proportion of each user profile according to the unemployment rate at the mobile phone tower level (Fig. S16), suggesting that unemployment rate does not influence the mobile phone behavior of users in Portugal.

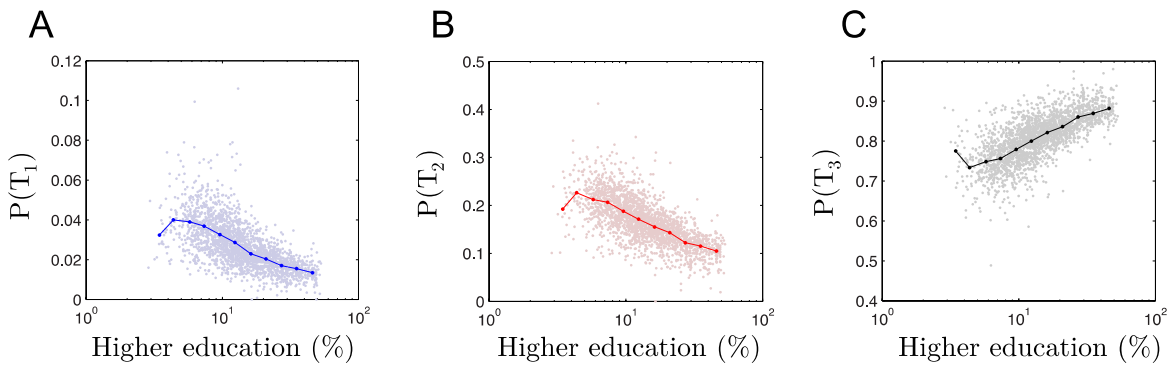

Figure S15: Variability of user profiles at each mobile phone tower according to the percentage of people holding a higher education degree. The proportion of (A) low and (B) medium activity users ($T_1$and T_2) tend to decrease with the education level, while the proportion of (C) high-activity users increases (T_3).

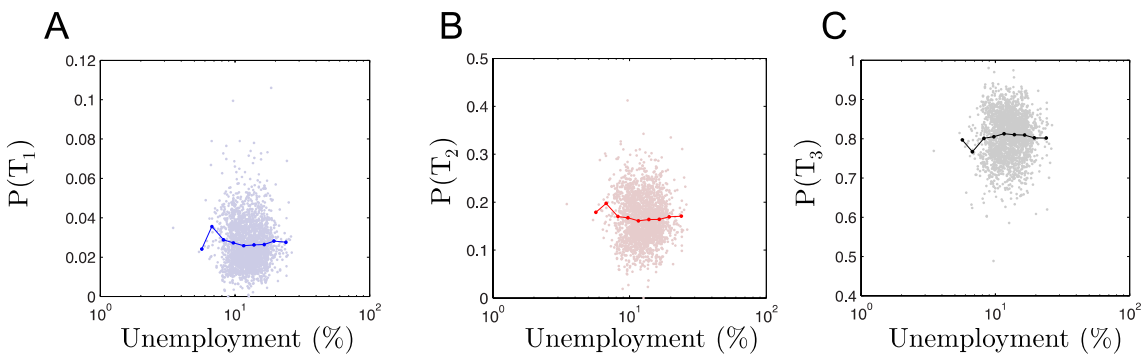

Figure S16: Variability of user profiles at each mobile phone tower according to the unemployment rate. We observe no correlation between unemployment rate and the proportion of (A) low, (B) medium and (C) high activity users.

C.4. Application to France

The population downscaling method developed in the present study was applied to France. Instead of the number of different users per night, we used here the number of daily-aggregated calls made or received from each tower during working periods (May, June, September, October 2007) for training the model. We have seen in section C.2. that using daily-aggregated call data had an impact on accuracy statistics, though this impact was largely due to the use of residential census data as reference for the accuracy assessment. The impact of using daily-aggregated call data on the estimation of β was rather low and not always significant.

We compared β coefficients calculated using the French dataset with the values we had for Portugal (using daily aggregated MP call data) in order to assess the variability of this coefficient between countries (Fig. S17). The standard and spatially-stratified cross-validation procedures defined in section B.1. were used to derive α and β coefficients for France. In order to use training datasets of comparable size for Portugal and France, only 2.5% of the 36,610 administrative units available for France were used as training data. Results show that β is higher in France (0.902 ±0.036) than Portugal (0.813 ±0.016) when estimated using a standard cross-validation procedure (Fig. S17A), but confidence intervals largely overlap when they are estimated using a spatially-stratified cross-validation procedure, with β values of 0.777 ±0.051 for Portugal and 0.846 ±0.056 for France (Fig. S17B). The larger confidence intervals observed for France are due to the higher number of administrative units available and the resulting greater diversity of administrative units sampled for training models.

In France, two regions (*Corse* and *Provence-Alpes-Cote-d'Azur*) are characterized by a particularly high proportion of tourists, with rates of camping area per person being the highest for these two regions (0.07 and 0.02 for the region of Corse and Provence-Alpes-Cote-d'Azur respectively, while the national average is 0.01) (13). When using these regions as training datasets, estimated β values are above 1, suggesting that a higher proportion of calls are made in low-density areas than in high-density areas in these regions. If we exclude these two regions from the training datasets, estimated β values are slightly lower (0.894 ±0.035 with a standard cross-validation procedure and 0.842 ±0.046 with a spatially-stratified cross-validation procedure). Choosing a training dataset that excludes the main holiday periods and typical tourism areas should thus be considered to reduce errors in population density estimates. It would indeed limit the discrepancies between residential and temporary population distributions.

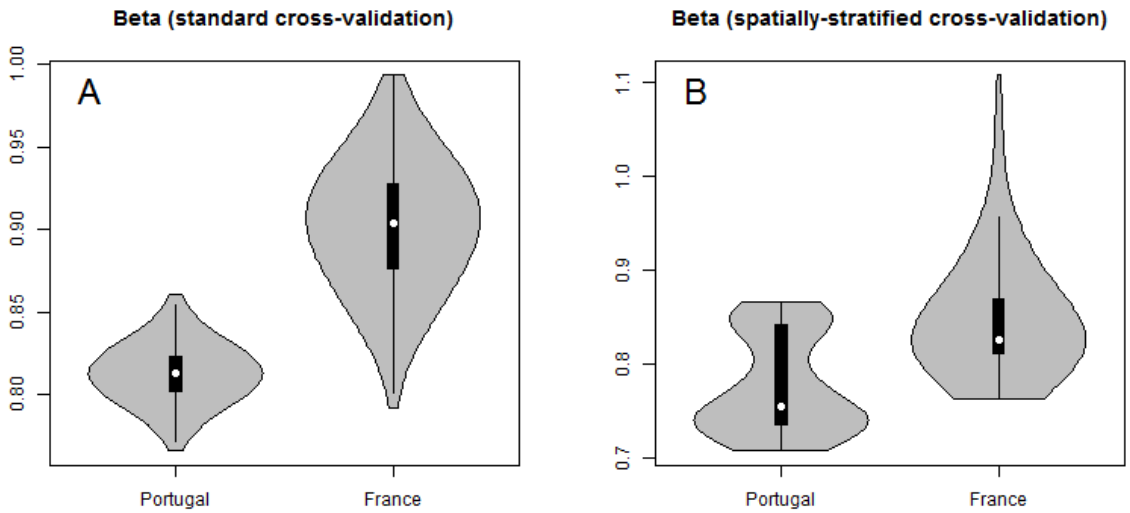

Figure S17: Comparison of β estimations in Portugal and France using (A) a standard cross-validation procedure and (B) a spatially-stratified cross-validation procedure.

D. Population dynamics

Temporal dynamics were derived from MP data using the timestamp associated to each MP call. Daily dynamics were analyzed by dividing the MP data into MP calls performed during the day (7 a.m. to 8 p.m.) and the night (8 p.m. to 7 a.m.). Weekly dynamics were analyzed by dividing the MP data into MP calls performed during weekdays (Monday to Friday) and MP calls performed during weekends (Saturday and Sunday). Seasonal dynamics were analyzed by dividing MP data into MP calls performed during the holiday period (July and August) and MP calls performed during working periods (all other months). Predicted population densities for each unit and for both time periods were computed using best-fit α and β estimates and relative differences between the two time periods were extracted.

The potential of MP data to estimate population density variations through time is illustrated in Fig. S18 for Portugal and Fig. S19 for France. Results show clear spatial patterns, such as population density increases along highways during the day (Fig. S18A), population density decreases in major cities during both weekends and holidays (Figs. S18B,C and S19) and important population density increases along the coast during holidays. Differences in estimated population densities between time periods are particularly important between day and night (Fig. S18A). These differences may be influenced by the variations in phone usage behaviors mentioned in section C.3. During the day, the proportion of low and medium-activity users is higher in densely populated areas, resulting in a lower number of phone calls per user. Such day/night variations are therefore more visible when using the number of users than the number of calls. This spatio-temporal variability in phone usage behaviors may influence population density estimates and emphasizes that, when data include users' identifiers, it is preferable to use the number of users than the number of calls. Some other phone usage behaviors may influence day/night variations such as the use of professional phones during the day and private phones during the night. Our results suggest that estimates may become more uncertain over shorter timescales.

We observed a positive correlation between the difference in estimated population between the holiday and the working periods and the number of tourist accommodations available by commune ($r = 0.28$, $p < 000.1$). The number of tourism accommodations (secondary residences and occasional accommodations, hotel rooms and camping plots) by commune in 2007 were downloaded from the INSEE website (www.insee.fr).

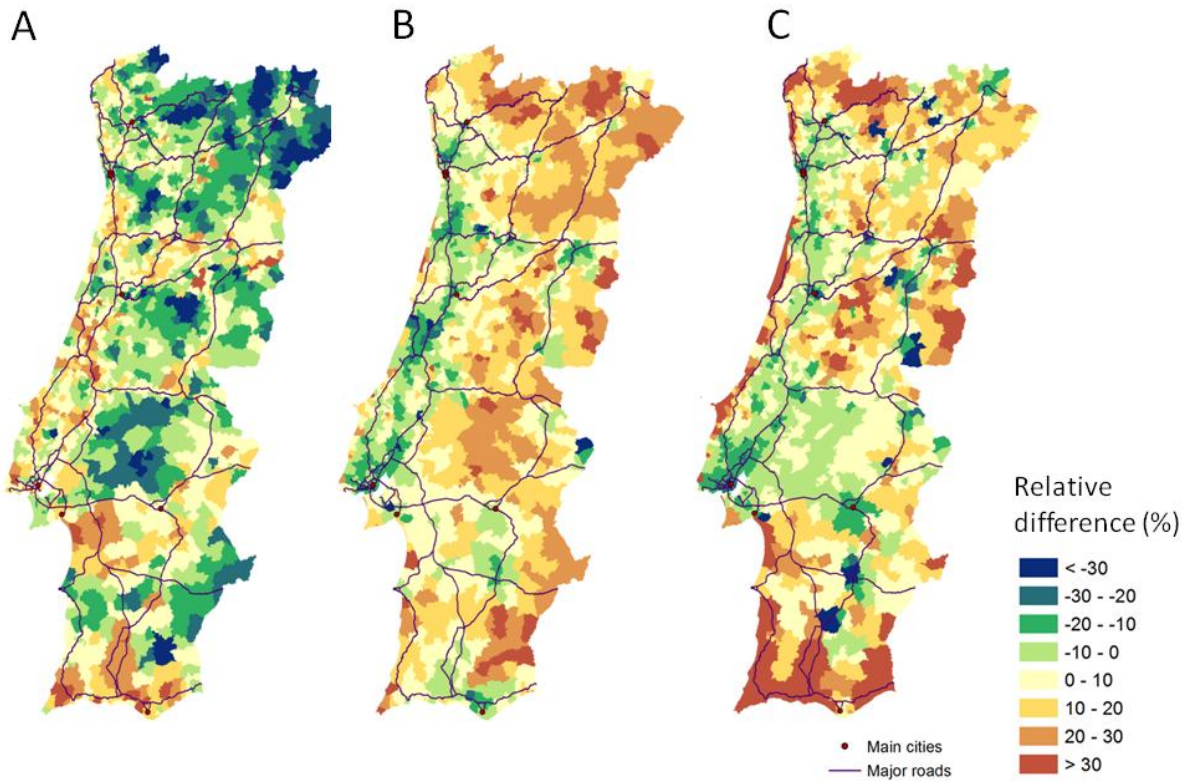

Figure S18: Relative difference in predicted population density by ADM-5 for different time periods in Portugal. (A) Difference between day and night, with brown colors indicating a higher population density during the day; (B) difference between weekend and weekdays, with brown colors indicating a higher population density during weekends; (C) difference between the main holiday period (July and August) and the working period (November-May), with brown colors indicating a higher population density during the holidays.

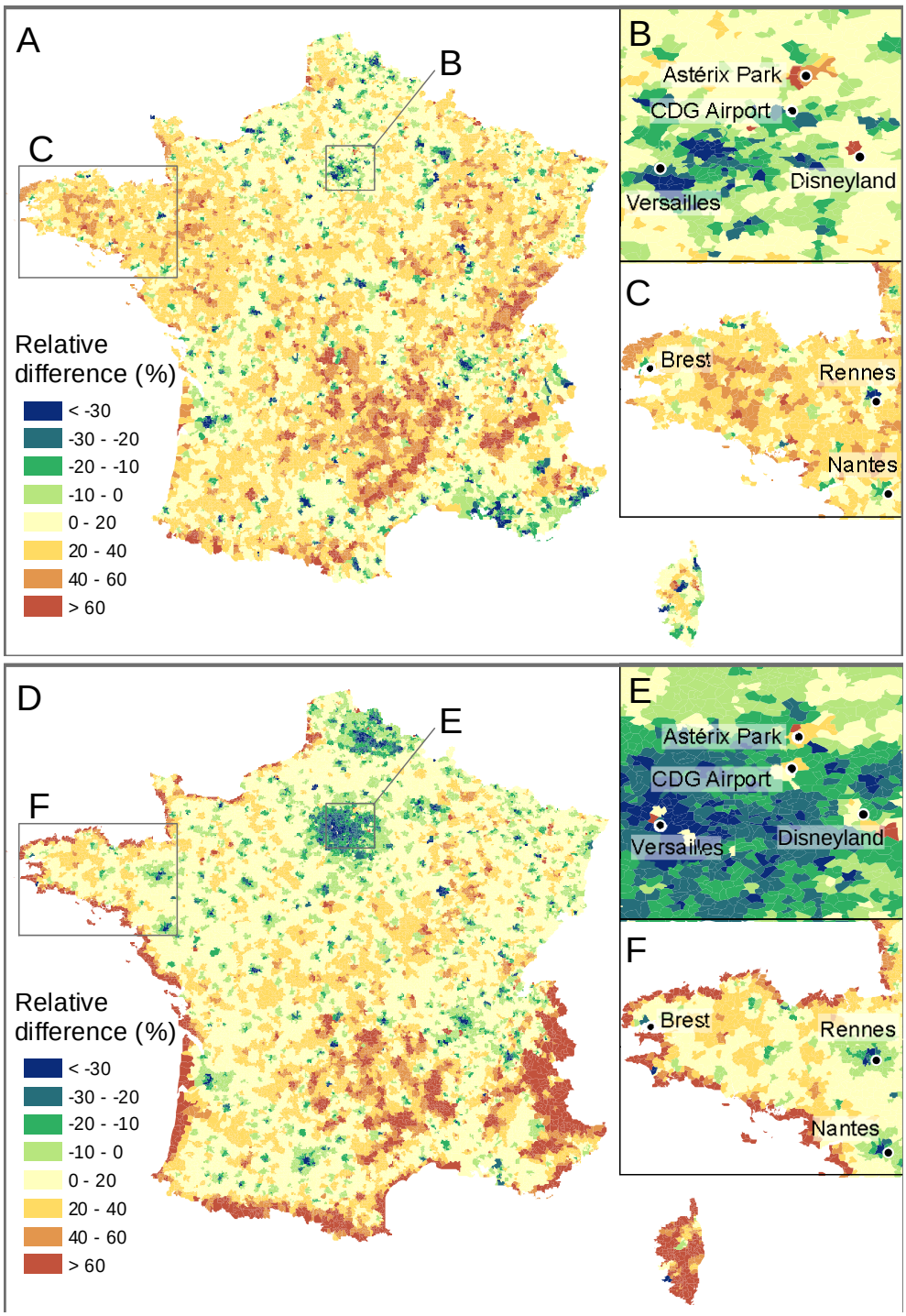

Figure S19: Relative difference in predicted population density by ADM-5 for different time periods in France. (A-C) Difference between weekend and weekdays. Brown colors indicate a higher population density during weekends. (D-F) Difference between the main holiday period (July and August) and the working period (May, June, September and October). Brown colors indicate a higher population density during holidays. (A,D) Metropolitan France; (B,E) close-ups around Paris; with labels showing the busiest airport in the country (Paris Charles de Gaulle), one of the most visited places in France (Palace of Versailles) and two popular recreation areas (Disneyland and Asterix Park) and (C,F) close-ups of the Bretagne Region, with labels showing the three most populated cities of the area: Rennes, Brest and Nantes.

References

1. Stevens FR, Gaughan AE, Linard C, Tatem AJ (In press) Disaggregating census data for population mapping using random forests with remotely-sensed and other ancillary data. *Plos One*.
2. Breiman L (2001) Random Forests. *Mach Learn* 45:5–32.
3. Liaw A, Wiener M (2002) Classification and Regression by randomForest. *R News* 2/3:18–22.
4. R. Core Team (2013) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria).
5. Autorité de Régulation des Communications Electroniques et des Postes (ARCEP). Available at: http://www.arcep.fr/ [Accessed February 2, 2014].
6. Bahn V, McGill BJ (2013) Testing the predictive performance of distribution models. *Oikos* 122:321–331.
7. Brenning A (2012) Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In *Geosci. Remote Sens. Symp. IGARSS 2012 IEEE Int.* 5372–5375. doi:10.1109/IGARSS.2012.6352393
8. Schläpfer M et al. (2014) The scaling of human interactions with city size. *J R Soc Interface* 11:20130789.
9. Gomez-Lievano A, Youn H, Bettencourt LMA (2012) The Statistics of Urban Scaling and Their Connection to Zipf's Law. *PLoS ONE* 7:e40393.
10. Krings G, Karsai M, Bernhardsson S, Blondel VD, Saramäki J (2012) Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Sci* 1:1–16.
11. Mitzenmacher M (2004) A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Math* 1:226–251.
12. Instituto Nacional de Estatística (2011) Censos 2011 - População residente por freguesia, CAOP 2013. Available at: www.ine.pt [Accessed January 30, 2014].
13. Institut National de la Statistique et des Etudes Economiques (2007) Population data from France. Available at: www.insee.fr [Accessed January 30, 2014].