Supplementary material

Table of Contents

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

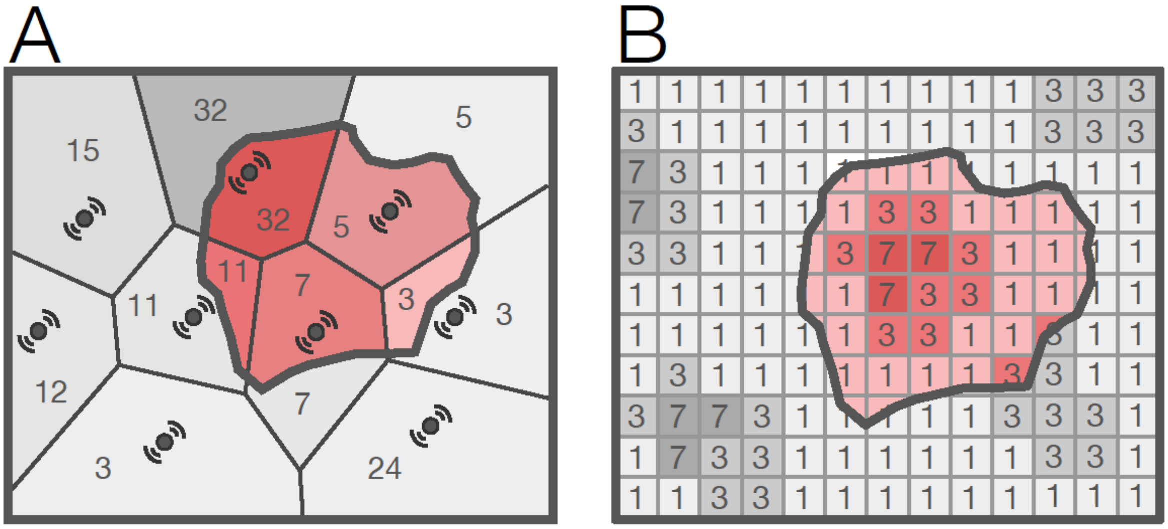

Figure S1: (A) Illustration of the *MP* method, where Voronoi polygons are built based on the spatial configuration of MP towers. The MP call density of an area (red polygon) is derived from the proportion of Voronoi polygons intersecting that area, as described in Equation 1, and the population density 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.

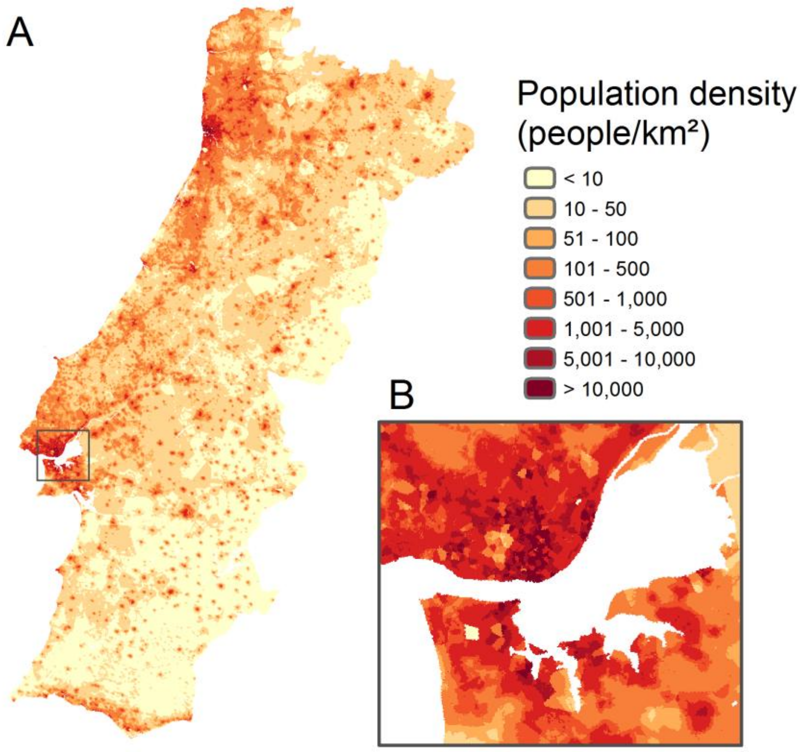

Figure S2: Population density at 100 x 100 m spatial resolution, as estimated by the combination of the *MP* and *RS* methods: (A) mainland Portugal with (B) close-up around the capital city Lisbon.

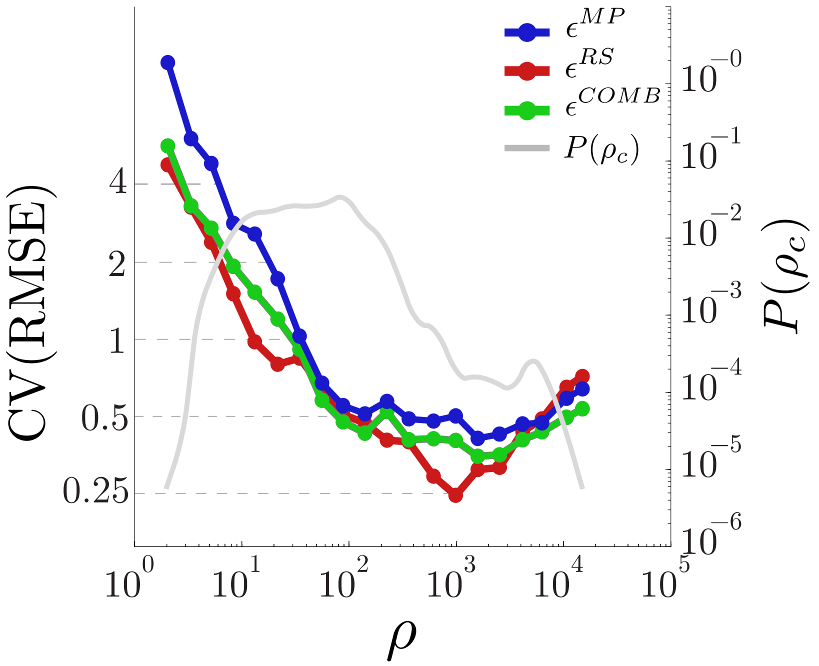

Figure S3: RMSEs normalized by the average population density of intervals, for the *MP* (blue), *RS* (red) and *COMB* methods (green). To aid visualisation, RMSEs are plotted on a logarithmic scale. The grey line represents the absolute population summed by population density intervals.

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.

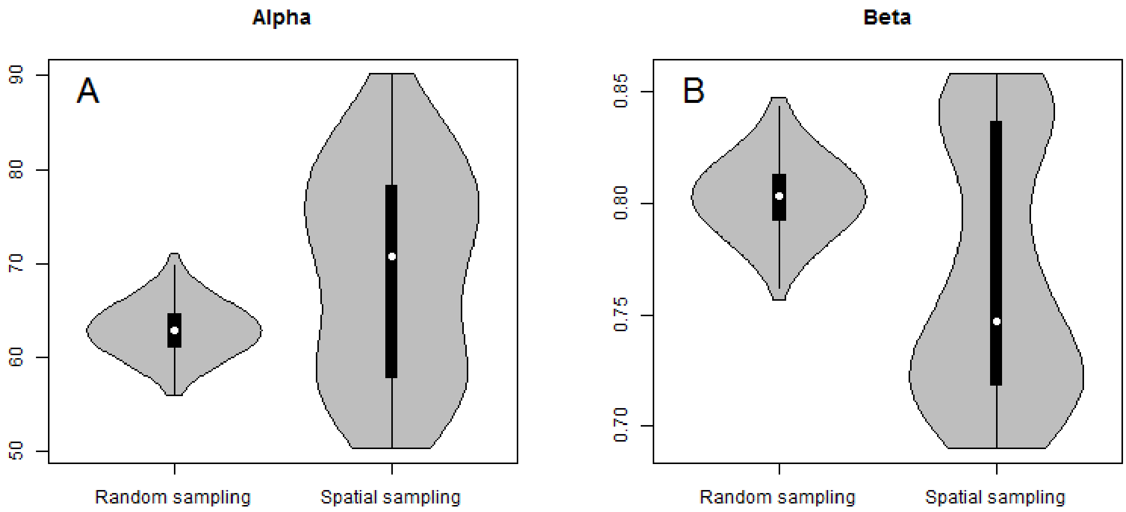

Figure S4: (A) Alpha and (B) beta coefficients estimated using randomly sampled and spatially-stratified training datasets.

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.

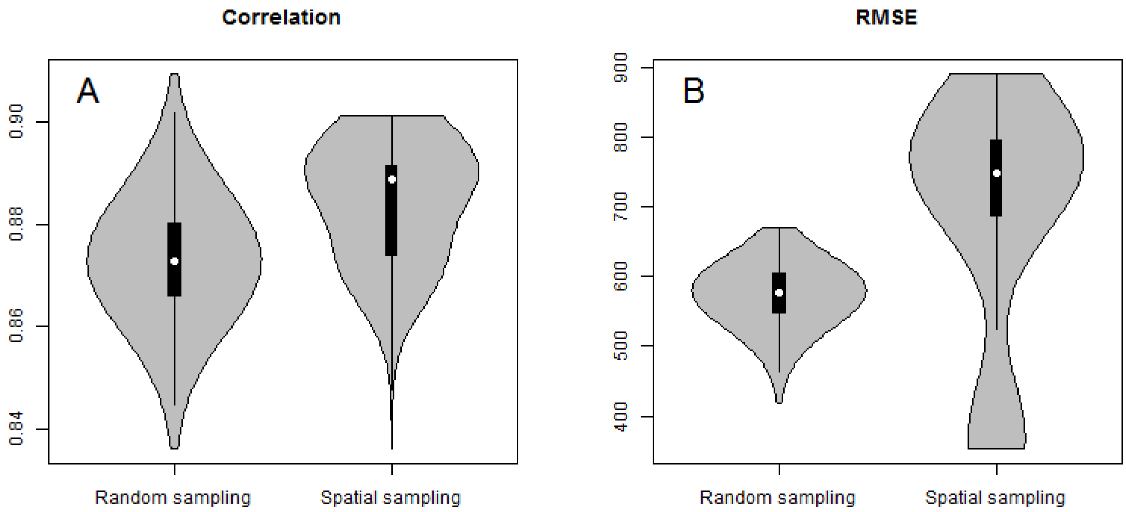

Figure S5: (A) Correlation coefficients and (B) RMSEs calculated using randomly sampled and spatially-stratified evaluation datasets.

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.

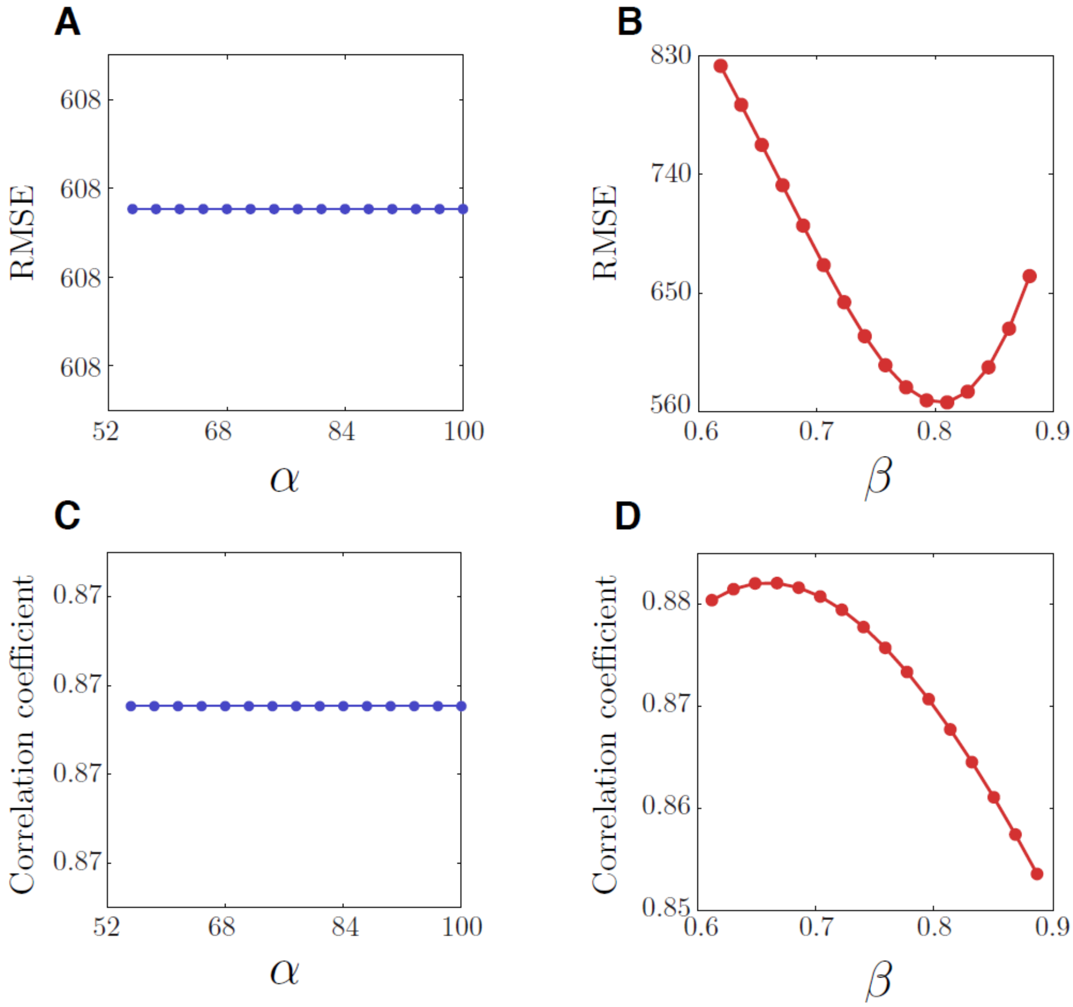

Figure S6: Influence of α and β parameters on the global RMSE and correlation coefficients.

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).

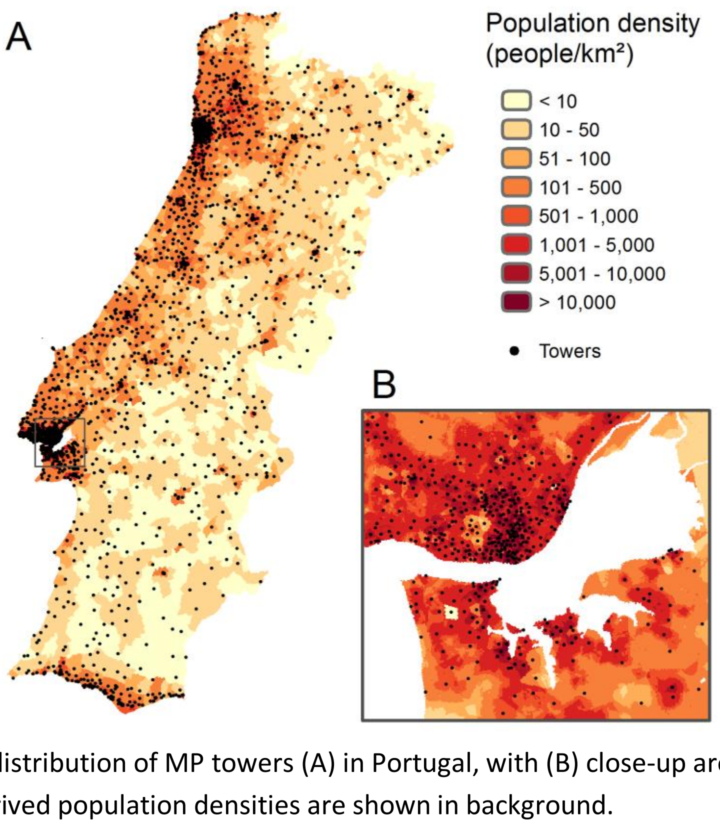

Figure S7: Spatial distribution of MP towers (A) in Portugal, with (B) close-up around the capital city Lisbon. Census-derived population densities are shown in background.

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.

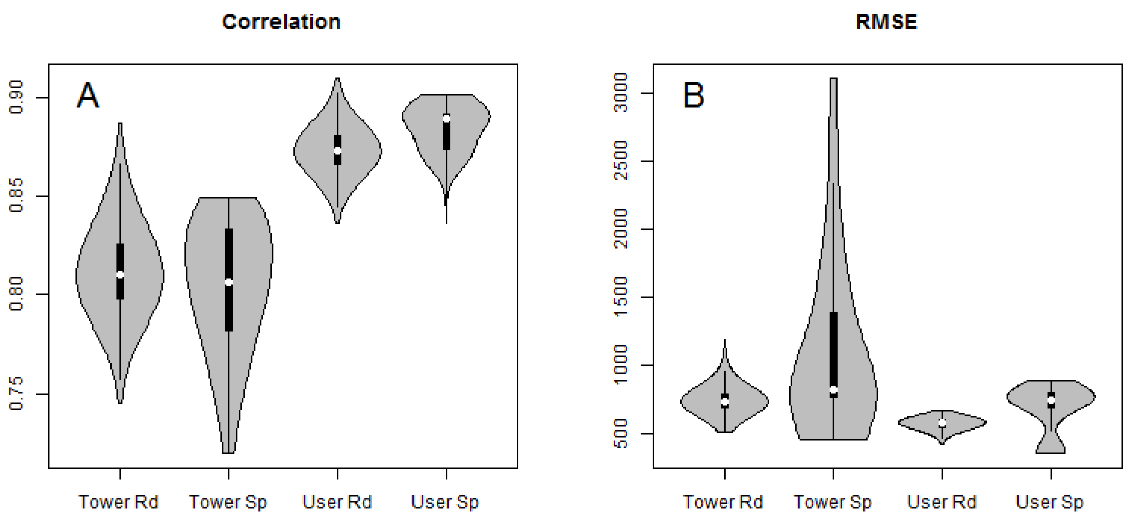

Figure S8: (A) Correlation coefficients and (B) RMSEs calculated using the density of phone towers and the density of users (Rd = standard cross-validation procedure; Sp = spatially-stratified cross-validation procedure)

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).

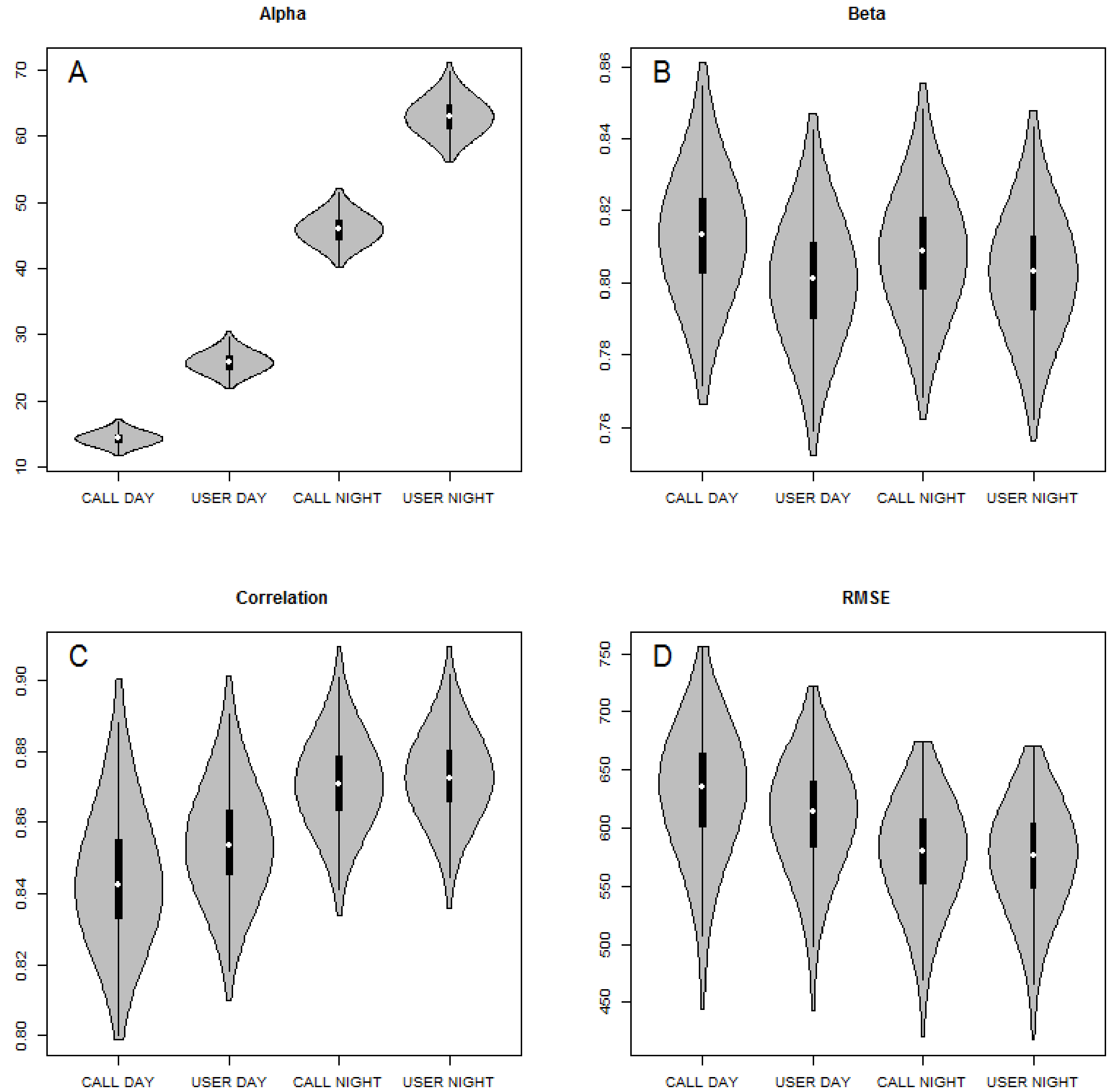

Figure S9: (A) Alpha, (B) beta, (C) correlation coefficient and (D) RMSE calculated when using (i) daily-aggregated calls (CALL DAY), (ii) daily-aggregated users (USER DAY), (iii) nighttime calls (CALL NIGHT) and (iv) nighttime users (USER NIGHT), with a standard cross-validation procedure.

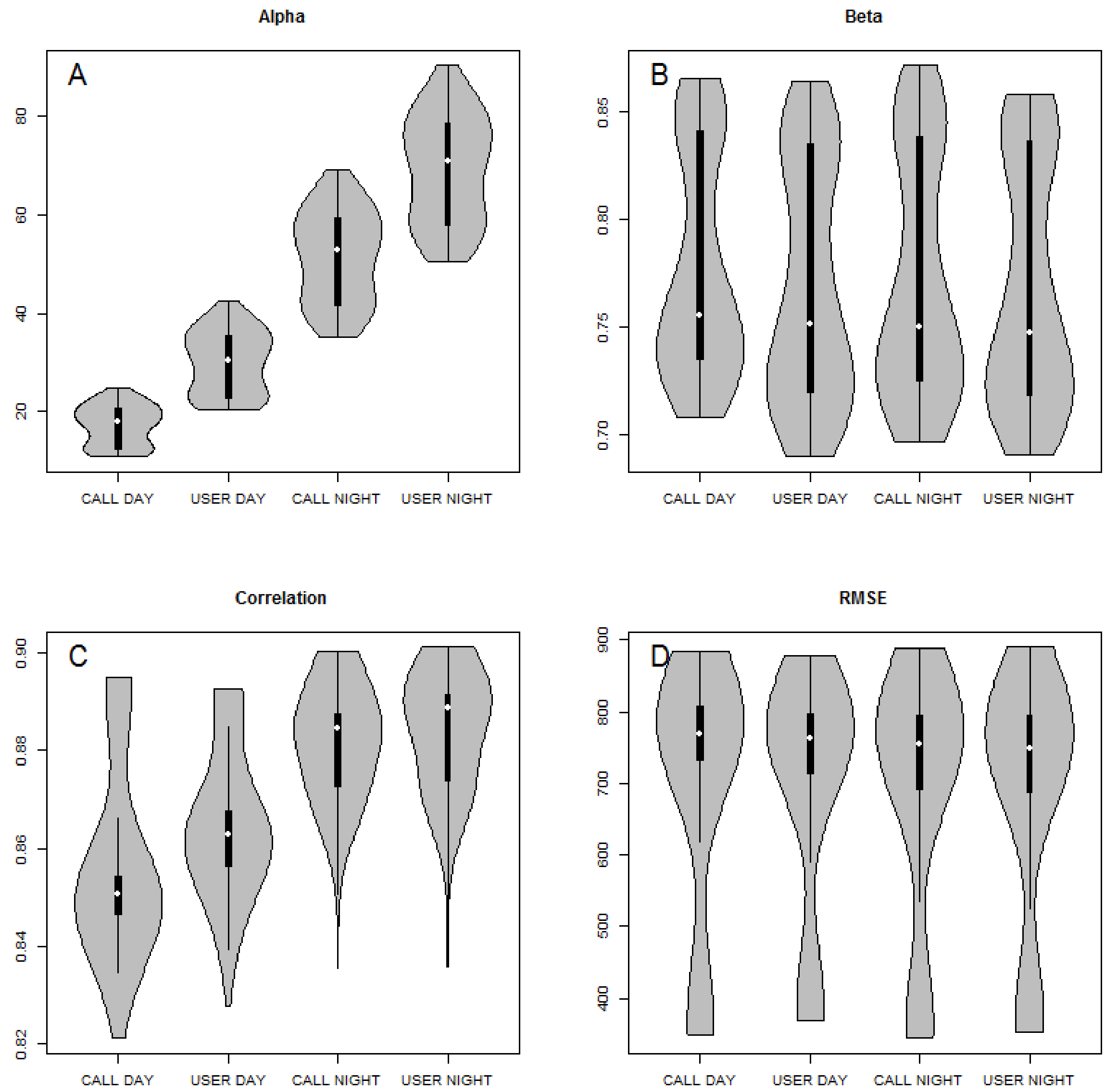

Figure S10: (A) Alpha, (B) beta, (C) correlation coefficient and (D) RMSE calculated when using (i) daily-aggregated calls (CALL DAY), (ii) daily-aggregated users (USER DAY), (iii) nighttime calls (CALL NIGHT) and (iv) nighttime users (USER NIGHT), with a spatially-stratified cross-validation procedure.

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).

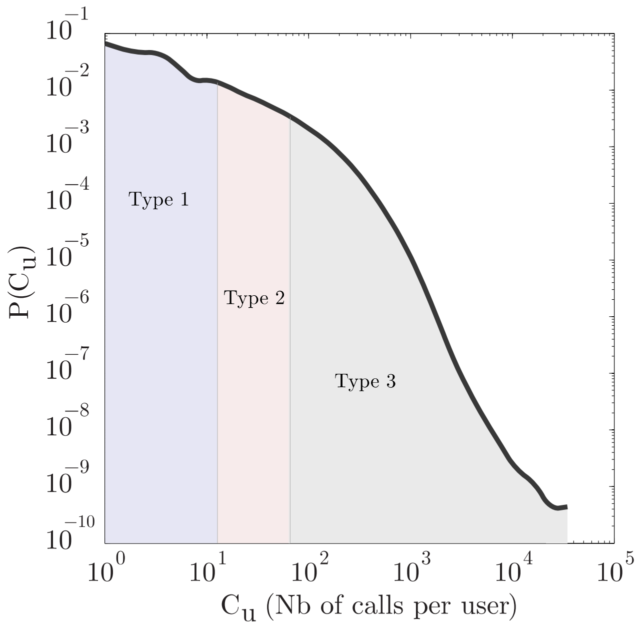

Figure S11: Probability Density Function of total number of night phone calls per user. Mobile phone users are divided into three distinct profiles, each containing a third of the users: low-activity users (Type 1), medium-activity users (Type 2) and high-activity users (Type 3).

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).

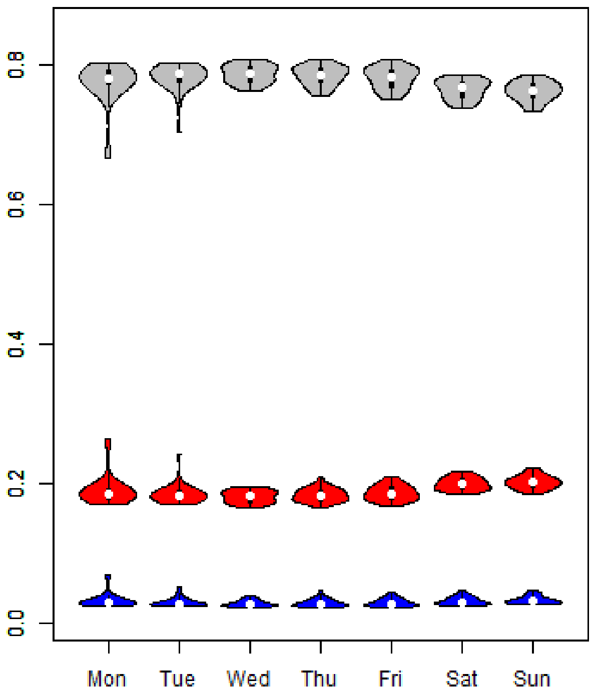

Figure S12: Variability of user profiles over time. Distribution of proportion of user of type 1 (blue), type 2 (red) and type 3 (grey) for each day of the week.

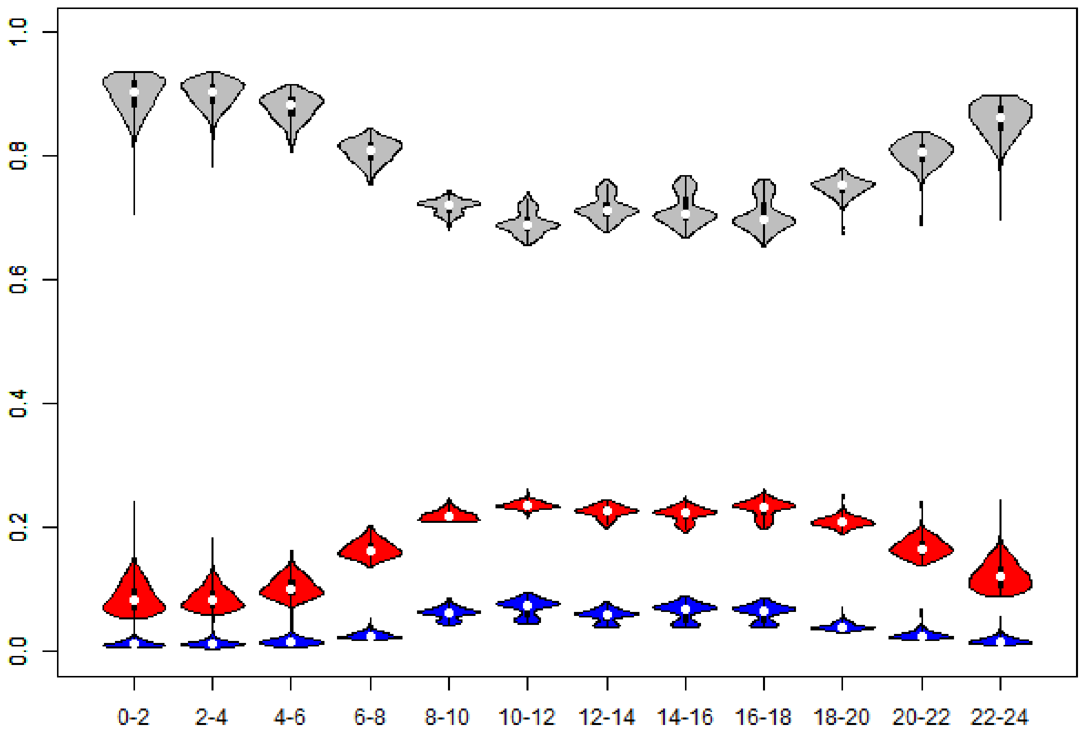

Figure S13: Variability of user profiles over time. Distribution of proportion of user of type 1 (blue), type 2 (red) and type 3 (grey) for each 2-hour period of the day.

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.

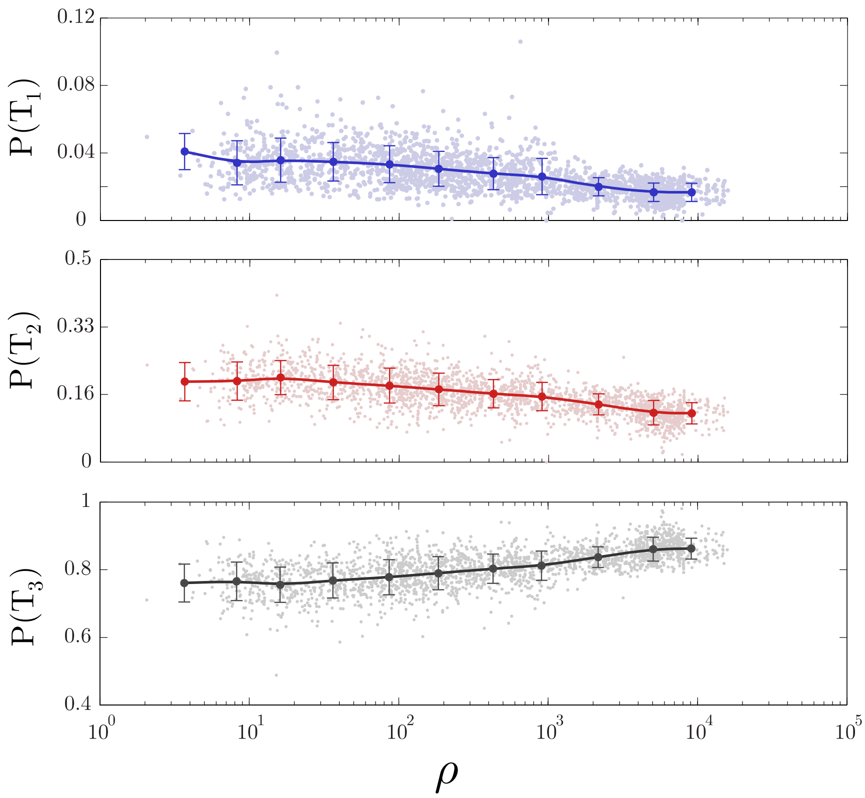

Figure S14: Variability of user profiles at each mobile phone tower over population density. The proportion of low (blue) and medium (red) activity users ($T_1$and T_2) tend to decrease in densely populated areas, while the proportion of high-activity users (grey) increases (T_3).

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.

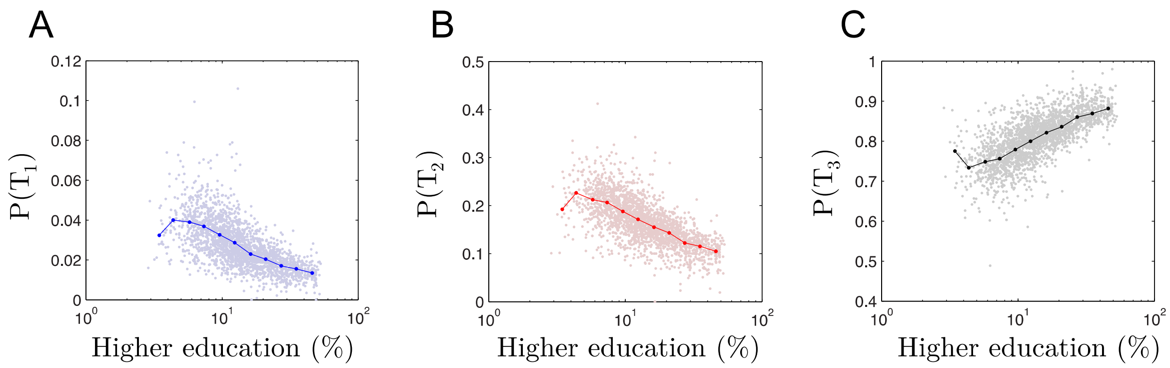

Figure S15: Variability of user profiles at each mobile phone tower according to the percentage of people holding a higher education degree. The proportion of (A) low and (B) medium activity users ($T_1$and T_2) tend to decrease with the education level, while the proportion of (C) high-activity users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

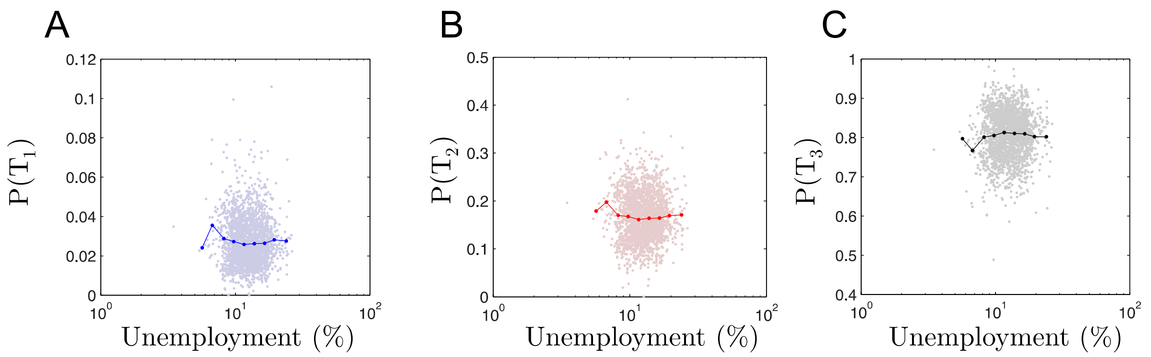

Figure S16: Variability of user profiles at each mobile phone tower according to the unemployment rate. We observe no correlation between unemployment rate and the proportion of (A) low, (B) medium and (C) high activity users.

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.

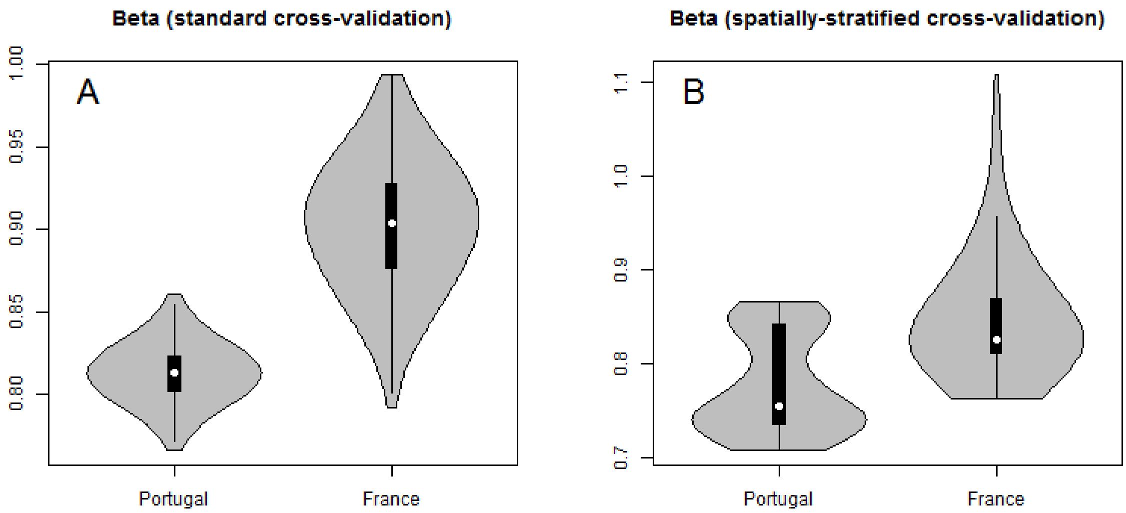

Figure S17: Comparison of β estimations in Portugal and France using (A) a standard cross-validation procedure and (B) a spatially-stratified cross-validation procedure.

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).

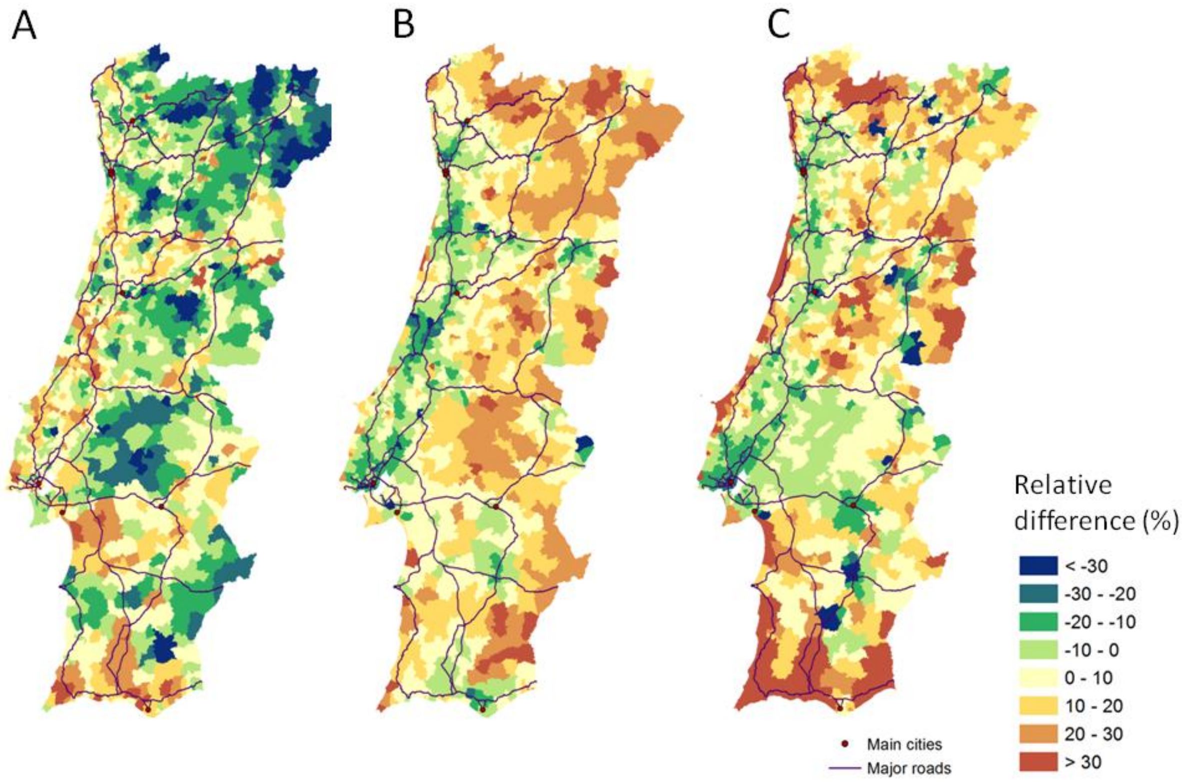

Figure S18: Relative difference in predicted population density by ADM-5 for different time periods in Portugal. (A) Difference between day and night, with brown colors indicating a higher population density during the day; (B) difference between weekend and weekdays, with brown colors indicating a higher 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.

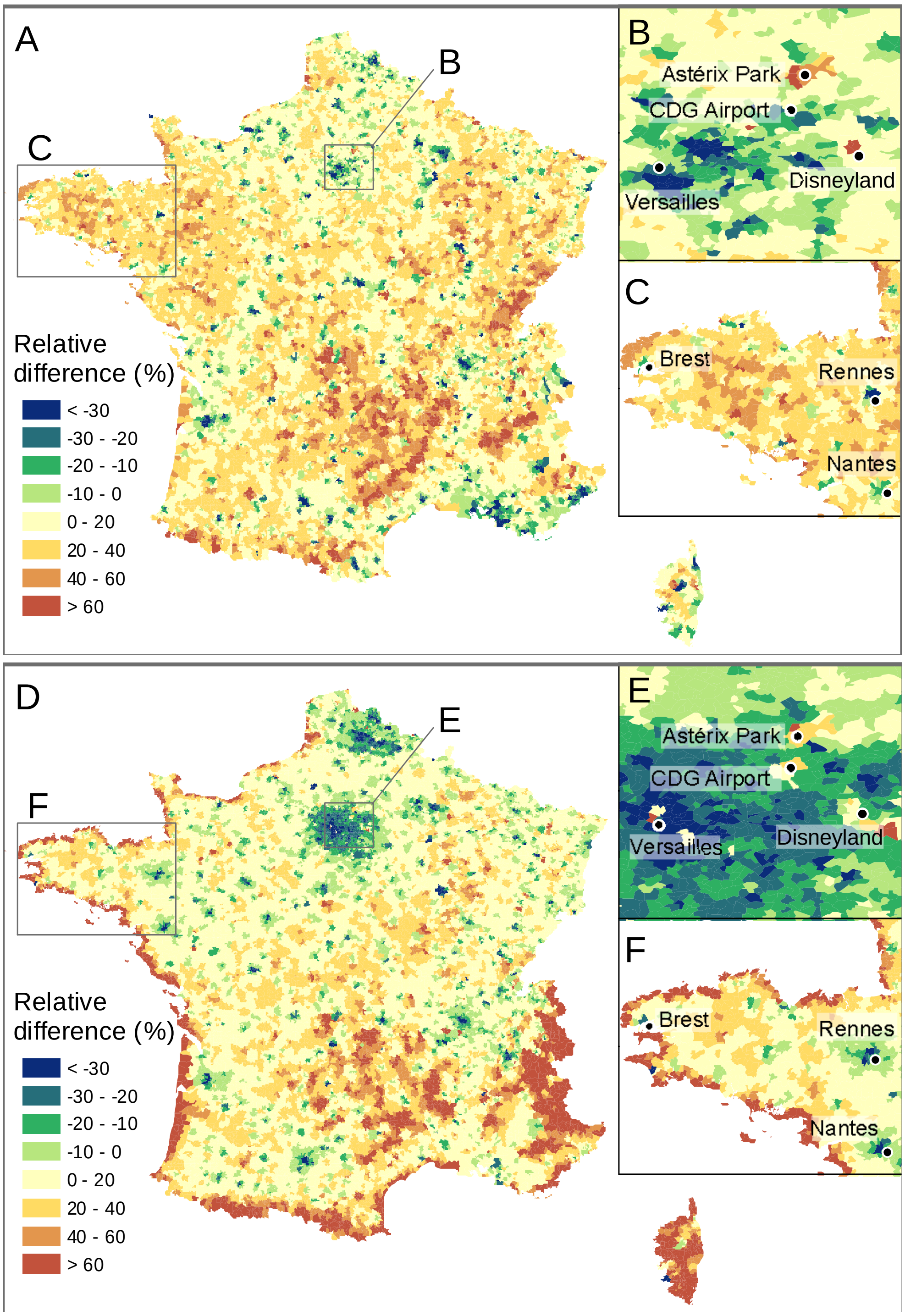

Figure S19: Relative difference in predicted population density by ADM-5 for different time periods in France. (A-C) Difference between weekend and weekdays. Brown colors indicate a higher population density during weekends. (D-F) Difference between the main holiday period (July and August) and the working 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].