High Resolution Population Mapping Using Spatial Point Process Models

Warren C. Jochem* and Andrew J. Tatem
School of Geography and Environmental Science
WorldPop Project
University of Southampton, UK
*w.c.jochem@soton.ac.uk

Abstract

Accurate and timely data on the population of local areas is vital for policy and decision making and for monitoring progress towards development goals. Yet in many places, population data are out of date and a complete population census is difficult to complete. This work addresses the challenge of producing accurate, high spatial resolution population estimates in the absence of a full census. We develop a marked spatial point process model to jointly model the density of building locations and population per building from samples of georeferenced households. We use a Bayesian framework and make predictions of the population with uncertainty at a 1 km grid cell resolution. We apply our model to a simulated georeferenced census which enables us to test different data sampling scenarios and evaluate predictions against a known population. The initial results suggest that point process models with a shared spatial effect have the potential to support population mapping and estimation; however more work is needed to investigate the sensitivity of the model. Ongoing work is also extending the basic joint model form to include geospatial ancillary data as covariates to improve the predictive performance. Further development of spatial point process models and related statistical techniques can open up opportunities to make fuller use of a wider range of datasets to study population distributions and to make accurate predictions of the population.

1. Introduction

Accurate and up-to-date information on the population and sociodemographic characteristics in local areas is vital for planning, carrying out, and evaluating health and development projects. Population data form the "denominator" used to calculate rates of disease, populations at risk, and many of the indicators for the Sustainable Development Goals (SDGs). In order to meet the SDG targets of leaving no one behind, more disaggregated datasets are required (Hosseinpoor, Bergen, & Magar, 2015), including subnational population estimates (Tatem, 2014). National censuses along with civil registration and vital records systems are fundamental sources of local population data, yet these data sources often remain tied to administrative unit boundaries and the finest spatial scale data are not made available due to privacy concerns. In order to address the need for more local-scale population data, geographers, spatial demographers, and other researchers have explored disaggregation techniques. These methods generally use geographic information system (GIS) techniques to allocate population to smaller area units or regularly sized grid squares. Well-known examples of gridded population datasets include WorldPop[1], LandScan[2], and

[1] https://www.worldpop.org/
[2] https://landscan.ornl.gov/

Gridded Population of the World[3]. For a recent review and comparison of gridded population datasets see Leyk et al. (2019).

In many places where population data are most needed for development aims, a census is too outdated to be reliably projected or a new census cannot be completed due to inaccessible areas, finances, and other challenges. For example, Afghanistan has not completed a census since 1979 and the Demographic Republic of the Congo's last census was in 1984. Disaggregation techniques to produce gridded population datasets rely on a census count or projection as input and therefore have limited accuracy in these situations. In the absence of a complete national census there have been several attempts recently to produce model-based population estimates using samples of observed population data (Weber et al., 2018). Wardrop et al. (2018) describe their approaches as "bottom-up" methods to population estimation. They use population counts within defined small areas which are statistically modelled with geospatial data, such as derived from satellites, to predict population in unobserved areas.

These population models show potential but they face several challenges. First, using aggregate population samples is often a necessary simplification for the statistical model or a limitation of data collection or processing, but this step obscures potentially important information on where households are located and how they are situated. Variations in settlement patterns can be indicative of neighbourhood types (Jochem, Bird, & Tatem, 2018) and can be related to differences in population density (Weber et al., 2018) or demographic rates (Benza, Weeks, Stow, López-Carr, & Clarke, 2017). Second, these models have so far relied heavily on having maps of settled or built-up areas in order to predict the number of inhabitants.

The present work addresses the challenge of accurate estimation and high-resolution mapping of population using bottom-up approaches in the absence of a full census, but in contrast to other models based on areal unit population counts, we develop a method for using georeferenced household-level data. Advances in GPS technology and a push for collecting georeferenced data during surveys and censuses are making spatially-explicit datasets more common. We develop a marked spatial point process model that jointly estimates the density of settlement locations along with household size using a shared spatial effect. Our approach focuses on using limited samples of population data and making predictions on high spatial resolution grids for unobserved areas. Here we demonstrate our method in a simulated census dataset that allows for a full evaluation of the predictions, but we will expand these analyses to real-world data where available.

2. Methods

2.1 Data

The data used are a synthetic, georeferenced census for the Oshikoto region of northern Namibia (Figure 1). Oshikoto is a primarily rural region with a mix of sparse settlements in the south and more densely populated in the western area. For these analyses we aggregate the number of household members to point locations of dwellings. Some households with the

[3] https://sedac.ciesin.columbia.edu/data/collection/gpw-v4

same spatial location (e.g. apartment buildings) are combined into a total count of people per building. A large area in the southwestern Oshikoto is a national park and we exclude this mostly uninhabited area for computational efficiency. The final dataset includes 36725 household locations with 178364 total people (people per household: Mean=4.9, SD=3.5).

The synthetic population data were produced using a combination of census microdata, household surveys, and digitised structure locations in order to produce a simulated point-level dataset with realistic distributions of characteristics. The methods and datasets are described in detail and made openly available in Thomson, Kools, and Jochem (2018). By using a synthetic dataset we are able evaluate the predictive performance of our models against a known, complete population.

2.2 Sampling

One application area of interest for the model developed here is to estimate and map populations when a census is geographically incomplete or to create intercensal estimates at a high spatial resolution from sparsely sampled data. Therefore we design our simulation study to only partially observe the population within Oshikoto while making predictions and evaluating the results for the entire study area. We present preliminary results from two simple data scenarios. The first scenario randomly samples 40 locations within Oshikoto. These points are buffered a random radius of 10 to 12 km, and any intersecting circles are merged to create the observation areas. The second sampling test uses a single, large observation area centred over the southern part of Oshikoto. The two sample designs are shown in Figure 1 and are referred to throughout as design “A” and “B” respectively. Data from the first design (A) might arise in the context of household surveys or vaccination campaigns which record georeferenced information about a household to track fieldwork. These unconventional data could be used to support population models, but are currently not utilised. The second sampling design (B) represents the scenario of a geographically incomplete census where an area was inaccessible to enumerators, but a full population estimate is needed.

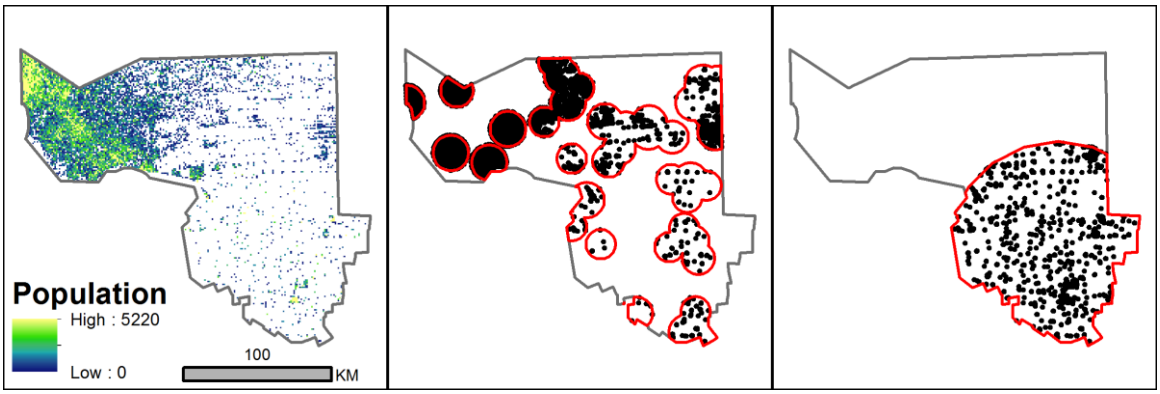

Figure 1: The study area of Oshikoto, Namibia. The simulated population is shown gridded to a 1km spatial resolution (left). Samples of household locations were taken in two forms for analysis: from patches randomly sampled across the area (design "A," shown middle) or from a concentrated area in the south (design “B,” shown right). The sample design figures include point locations of observed households.

2.3 Marked spatial point process model

In conceptual terms, the total population in a given location is related to the number of households in an area and the number of people in each household. In places with more

dwellings per area we might expect more people, but a higher density of dwellings may also suggest an urban area with different family structures and smaller average household sizes, for example. By modelling how the density of house locations varies in space jointly with the variation in population size we are able to include such a relationship. Moreover, the joint model allows us to estimate the population without complete observations of all households or a complete map of settlement locations in the study area. In this work, the spatial patterns of the residential buildings in the study are represented as a spatial point process. Spatial point process methods have been discussed in depth previously (Illian & Burslem, 2017; Illian, Sørbye, & Rue, 2012; Møller, Syversveen, & Waagepetersen, 1998). Briefly, a spatial point process describes the random patterns of observed events in space, the locations of which depend on underlying spatial processes. In the case of housing, these may relate to human-environment interactions and political economic factors that influence where people settle. The number of points within some area, A, is a random variable modelled by an intensity function, $\lambda(s)$, such that $N(A) = \int_A \lambda(s)\, ds$. A relevant model for these counts is a Poisson process with an intensity function that allows for different spatial patterns and does not assume complete spatial randomness. We use the commonly applied technique of a log-Gaussian Cox process (LGCP)(Møller et al., 1998). In a LGCP the intensity is itself stochastic and is allowed to vary over space, $\lambda(s), s \in S$. The log intensity can then be modelled with a Gaussian linear predictor that can be easily extended to include covariates and other effects such as a spatially structured random field.

While the LGCP is used to model the point patterns, each location contains an additional observed quantity, $v(s)$, knowns as the "marks" which in our study is the population total. The marks in our case are always positive values greater than zero (i.e. no abandoned houses are observed) and are assumed to follow a lognormal distribution. The two components have separate likelihoods but are modelled jointly as seen in the following. For the points, the intensity of the LGCP is modelled as:

$$log\{\lambda(s)\} = \alpha_{01} + \sum \beta_{i1} z_i(s) + W_s,$$

and the marks as:

$$v(s) = \alpha_{02} + \sum \beta_{j2} z_j(s) + \beta_s W_s,\ s \epsilon S$$

In this format, α_{01} and α_{02} are intercepts, β_{i1} and β_{j2} are model parameters to be estimated for any covariates, z_i and z_j respectively. W_s is a latent, zero-mean Gaussian Markov field with Matérn covariance structure. This spatial field is shared between the two model components, setting up the possibility of dependence between the point pattern and marks. Because the marks and points are on different response scales, the spatial field for the marks is scaled by the estimated parameter β_s.

The spatial effect, $W_s, s \epsilon S$, deserves more attention. While past point process models use a regular grid (or lattice) structure to aggregate and model the counts, Simpson, Illian, Lindgren, Sørbye, and Rue (2016) demonstrated defining the spatial effect in continuous space which makes use of the exact point locations. To model a spatial effect in continuous space we use the spatial partial differential equation (SPDE) approach (Lindgren, Rue, & Lindström, 2011) to approximate a Gaussian field (GF) with Mátern covariance structure. The SPDE approach uses a set of basis functions on a triangular mesh to represent a GF as a

Gaussian Markov random field (GMRF) which offers flexibility and computational advantages.

We implemented our marked spatial point process model in a Bayesian framework using the integrated nested Laplace approximation (INLA) approach implemented in R-INLA (Rue, Martino, & Chopin, 2009) and with the R package inlabru (Bachl, Lindgren, Borchers, Illian, & Freckleton, 2019). Following Fuglstad, Simpson, Lindgren, and Rue (2018) we constructed a penalised complexity prior for the range (ρ) and standard deviation (σ) of the spatial field. This method specifies the hyperparameters such that $P(\rho < \rho_0) = \alpha_1$ and $P(\sigma > \sigma_0) = \alpha_2$. We used $\rho_0 = 100$, $\sigma_0 = 5$, $\alpha_1 = 0.5$, $\alpha_2 = 0.5$, and we assigned Gaussian prior distributions with mean=0 and precision=1E-3 for fixed effects.

3. Preliminary results

For this preliminary study we present results from models which include intercepts and a spatial field but not covariates. The main outcome of interest of our model is the predicted population for both the total study region and at local scales. To make spatial predictions we constructed a regular grid with 1km x 1km grid cells covering the study area. The predicted population in each location of the grid is estimated by sampling from the posterior of the joint model. We express the predicted population as the mean of the posterior samples and the uncertainty around that prediction as the 95% confidence intervals. To evaluate the predictions at the grid cell level we convert the point locations and population of the synthetic data to a grid with matching 1km resolution.

Sample design A resulted in observations of 12347 locations (approximately 34% of all locations) totalling 59777 people in 16 areas across the study area (Figure 1). The spatial range for the model using data from design A was estimated to be 73.7 km (95% CI: 50.8 to 110.1) with a standard deviation of 2.03 (95% CI: 1.5 to 2.9). The results of the predicted population grid are shown in Figure 2 along with the upper and lower bounds of the confidence interval. The total population was predicted to be 182122 (95% CI: 154436 to 242798) which is slightly higher than the true population of 178364 people. A cell-level comparison of the true population and mean prediction is shown in Figure 4 showing areas of both over- and under-prediction.

Sample design B covered a single large area in the southern portion of Oshikoto containing 7270 point locations (approximately 20% of the locations) and totalling 28809 people (Figure 1). The spatial range for the model using data from design B was estimated to be 20.9 km (95% CI: 15.5 to 28.0) with a standard deviation of 1.27 (95% CI: 1.07 to 1.51). The results of the predicted population grid are shown in Figure 3 along with the upper and lower bounds of the confidence interval. The total population was predicted to be 43982 (95% CI: 36533 to 54330). A cell-level comparison of the true population and mean prediction is shown in Figure 4 showing a large area in the west of the study area, far from the sample domain that was substantially underpredicted.

Figure 2: Predicted population in 1km resolution grids from data scenario A. The mean predicted population (left) along with the lower bound (middle) and upper bound (right) of the 95% confidence interval.

Figure 3: Predicted population in 1km resolution grids from data scenario B. The mean predicted population (left) along with the lower bound (middle) and upper bound (right) of the 95% confidence interval.

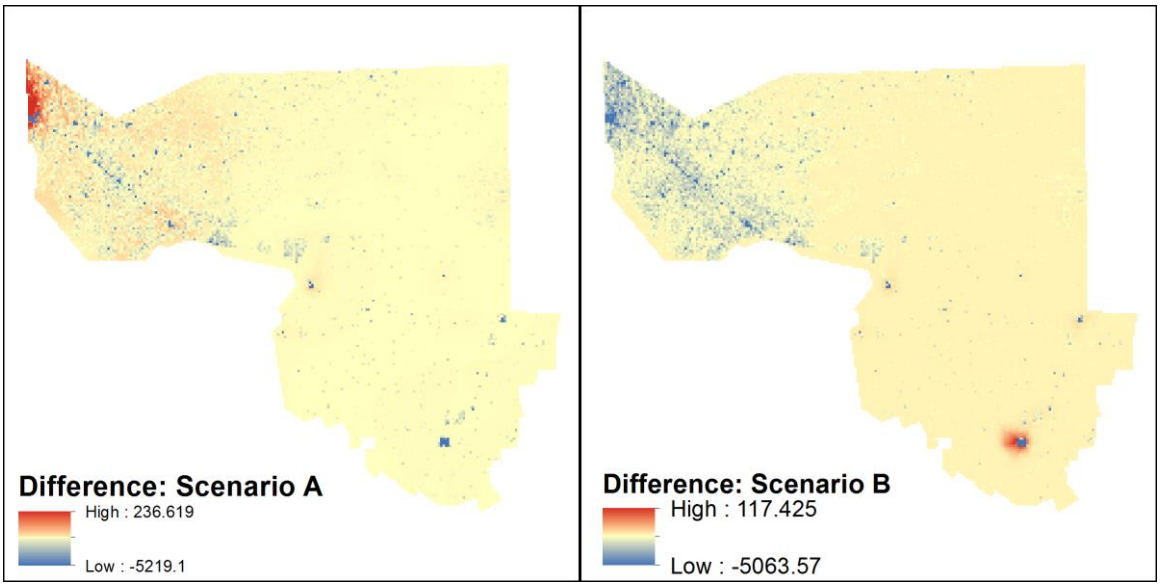

Figure 4: Comparison between the predicted and true populations at the 1km cell level. The difference in mean predicted population and true population are shown for scenario A (left) and scenario B (right). Positive values indicate an overprediction while negative values are an underprediction.

4. Preliminary conclusions

Marked spatial point process models have been applied most commonly in ecology to study the abundance of wildlife or environmental resources. Spatial demographers and population geographers have not commonly used these models, likely owing to limited availability of georeferenced point-level data. However, one notable exception in this area is work by (Pereira, Turkman, Correia, & Rue, 2019) who used a marked point process model to estimate the unemployed population in Portugal from georeferenced household surveys. We have demonstrated preliminary results from a joint spatial model of settlement point pattern and population. We applied our model to samples of building-level data from a synthetic census population which allowed us to evaluate the predicted population for the total study area and at the local scale (aggregated within 1km grid cells).

The model was efficient to fit using INLA and SPDE methods, and the result show promise despite using only a sample of the location data and given that no covariate data were used in fitting. The model fit with data observed from across the study region (design A) performed best and correctly predicted closely the total population of the study region. At the grid cell level it tended to underestimate the highest population areas while slightly overestimating primarily in the western regions (Figure 4). This result suggests the model is oversmoothing and not predicting very small scale variation population. When data were sampled from only one part of the study region (design B), the result was a significant underprediction of the total population in the region. We note that the sample came from the southern area which has a lower population density, thus biasing the sample. In the absence of other ancillary data to assist the model predictions, the poor predictive performance was unsurprising for sample B.

Overall our spatial modelling approach provides a flexible framework for modelling population distribution and making predictions with uncertainty. It is clear from these initial results that future work should seek to refine the model of household size and to incorporate ancillary data as covariates to help explain the variation in building locations and thus improve predictions. We also plan to further explore the sensitivity of sample design and the effect on the accuracy of the predictions. Further development of spatial point process models and related statistical techniques can open up opportunities to make fuller use of a wider range of datasets to study population distributions and to make more accurate predictions of the population in local areas.

References

Bachl, F. E., Lindgren, F., Borchers, D. L., Illian, J. B., & Freckleton, R. (2019). inlabru: an R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution, 10*(6), 760-766. doi:10.1111/2041-210x.13168

Benza, M., Weeks, J. R., Stow, D. A., López-Carr, D., & Clarke, K. C. (2017). Fertility and urban context: A case study from Ghana, West Africa, using remotely sensed imagery and GIS. *Population, Space and Place, 23*(8), e2062. doi:10.1002/psp.2062

Fuglstad, G.-A., Simpson, D., Lindgren, F., & Rue, H. (2018). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association, 114*(525), 445-452. doi:10.1080/01621459.2017.1415907

Hosseinpoor, A. R., Bergen, N., & Magar, V. (2015). Monitoring inequality: an emerging priority for health post-2015. *Bull World Health Organ, 93*(9), 591-591A. doi:10.2471/BLT.15.162081

Illian, J. B., & Burslem, D. F. R. P. (2017). Improving the usability of spatial point process methodology: An interdisciplinary dialogue between statistics and ecology. *AStA Advances in Statistical Analysis, 101*(4), 495-520. doi:10.1007/s10182-017-0301-8

Illian, J. B., Sørbye, S. H., & Rue, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *The Annals of Applied Statistics, 6*(4), 1499-1530. doi:10.1214/11-aoas530

Jochem, W. C., Bird, T. J., & Tatem, A. J. (2018). Identifying residential neighbourhood types from settlement points in a machine learning approach. *Comput Environ Urban Syst, 69*, 104-113. doi:10.1016/j.compenvurbsys.2018.01.004

Leyk, S., Gaughan, A. E., Adamo, S. B., de Sherbinin, A., Balk, D., Freire, S., . . . Pesaresi, M. (2019). The spatial allocation of population: A review of large-scale gridded population data products and their fitness for use. *Earth System Science Data, 11*(3), 1385-1409. doi:10.5194/essd-11-1385-2019

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stohastic partial differential equation approach. *Journal of the Royal Statistical Society B: Statistical Methodology, 73*(Part 4), 423-498.

Møller, J., Syversveen, A. R., & Waagepetersen, R. P. (1998). Log Guassian Cox Processes. *Scandinavian Journal of Statistics, 25*, 451-482.

Pereira, S., Turkman, K. F., Correia, L., & Rue, H. (2019). Unemployment estimation: Spatial point referenced methods and models. *Spatial Statistics*, 100345. doi:10.1016/j.spasta.2019.01.004

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society B: Statistical Methodology, 71*(Part 2), 319-392.

Simpson, D., Illian, J. B., Lindgren, F., Sørbye, S. H., & Rue, H. (2016). Going off grid: computationally efficient inference for log-Gaussian Cox processes. *Biometrika, 103*(1), 49-70. doi:10.1093/biomet/asv064

Tatem, A. J. (2014). Mapping the denominator: Spatial demography in the measurement of progress. *International Health, 6*(3), 153-155. doi:10.1093/inthealth/ihu057

Thomson, D., Kools, L., & Jochem, W. (2018). Linking synthetic populations to household geolocations: A demonstration in Namibia. *Data, 3*(3), 30. doi:10.3390/data3030030

Wardrop, N. A., Jochem, W. C., Bird, T. J., Chamberlain, H. R., Clarke, D., Kerr, D., . . . Tatem, A. J. (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. *Proc Natl Acad Sci U S A, 115*(14), 3529-3537. doi:10.1073/pnas.1715305115

Weber, E. M., Seaman, V. Y., Stewart, R. N., Bird, T. J., Tatem, A. J., McKee, J. J., . . . Reith, A. E. (2018). Census-independent population mapping in northern Nigeria. *Remote Sens Environ, 204*, 786-798. doi:10.1016/j.rse.2017.09.024