WorldPop Book of Methods, Vol. I: Gridded Population Estimates

WorldPop, University of Southampton

2020-11-26

Contents

About this book

Notice: This book is still being drafted! All text is subject to change. Please do not redistribute it (i.e. the code, text, images, tables, etc.) until this notice has been removed.

This open book provide a guide to WorldPop's gridded population data sets and the methods used to create them. It was developed by the WorldPop Research Group within the Department of Geography and Environmental Science at the University of Southampton. Individual contributers, collaborators and funders are recognized within each chapter. Please refer to individual chapters for suggested citations or cite the whole book as:

> WorldPop. 2020. *WorldPop Book of Methods, Vol. I: Gridded Population Estimates.* WorldPop, University of Southampton. 26 November 2020. https://docs.worldpop.org.

The source code for the book is available from WorldPop on GitHub: https://github.com/wpgp/bookworm.

Introduction

This section will introduce you to WorldPop population data sets and various ways to access them.

Chapter 1

Gridded Population Estimates

> **Note:** This chapter is currently being drafted. To make suggestions, please raise an issue in the bookworm repository. To make direct edits to the source code, please submit a pull request to merge your code with the "dev" branch. Include the tag @doug-leasure in your issues or pull requests to notify Doug.

WorldPop produces population estimates for every 100 x 100 meter grid square with national coverage for countries throughout the world. Gridded data allow end-users to aggregate grid cells to estimate the total population within any boundaries that meet their needs. It also provides a consistent grid to facilitate combining data sets like population, age-sex structure, vaccinations, maternal health, poverty, etc.

There are three broad categories of methods that WorldPop uses to produce gridded population estimates:

1. Top-down
2. Bottom-up
3. Peanut butter

These approaches require different inputs (i.e. population data) and the gridded population estimates that they produce have different characteristics (Fig. **??**). Top-down methods require data from a complete recent census or use other admin totals in each administrative unit (or other geographic unit). Bottom-up methods require data from geolocated household surveys that contain a representative sample of locations across the country. For top-down approaches, the population totals for adminstrative units are pre-defined by the input data, whereas in bottom-up approaches, the population totals are an emergent property that is not pre-defined.

The peanut butter method is different because it relies solely on expert opinion (rather than hard data) to define the average number of people per building. The strength of this method lies in the quality of the high resolution building footprints that are used (**?**). This approach may be a good option when recent and reliable data from a national census and/or household surveys are not available.

1.1 Top-down

The top-down method (**?**) disaggregates known population totals for each administrative unit (e.g. states or local government areas) into 100 m gridded population estimates (Fig. **??**). Population totals may be obtained from national census results projected to the current year (or other years). Gridded population estimates are created by using machine learning methods (random forest models) to disaggregate population totals based on relationships with spatial covariates such as building density, distance to city center, or intensity of nighttime lights. The disaggregation can be applied across the entire country or constrained only to areas where settlements have been mapped.

This is a good approach when recent census totals are available. These models perform best when accurate census totals are available for the smallest administrative units. When census results are outdated, this method relies on projections that can introduce error. Top-down methods generally do not produce estimates of uncertainty.

See the Top-down Models section for details.

1.2 Bottom-up

The bottom-up method (**??**) uses geolocated household survey data from a sample of locations to fit statistical models that estimate population sizes for unsampled areas based on relationships with spatial covariates (Fig. **??**). This approach applies customized statistical models to make the best use of available survey data and to provide probabilistic estimates of uncertainty.

This is a good approach when recent geolocated household survey data are available where there has not been a recent or complete national census. This approach provides robust estimates of uncertainty but requires more detailed input data and more time to develop the models.

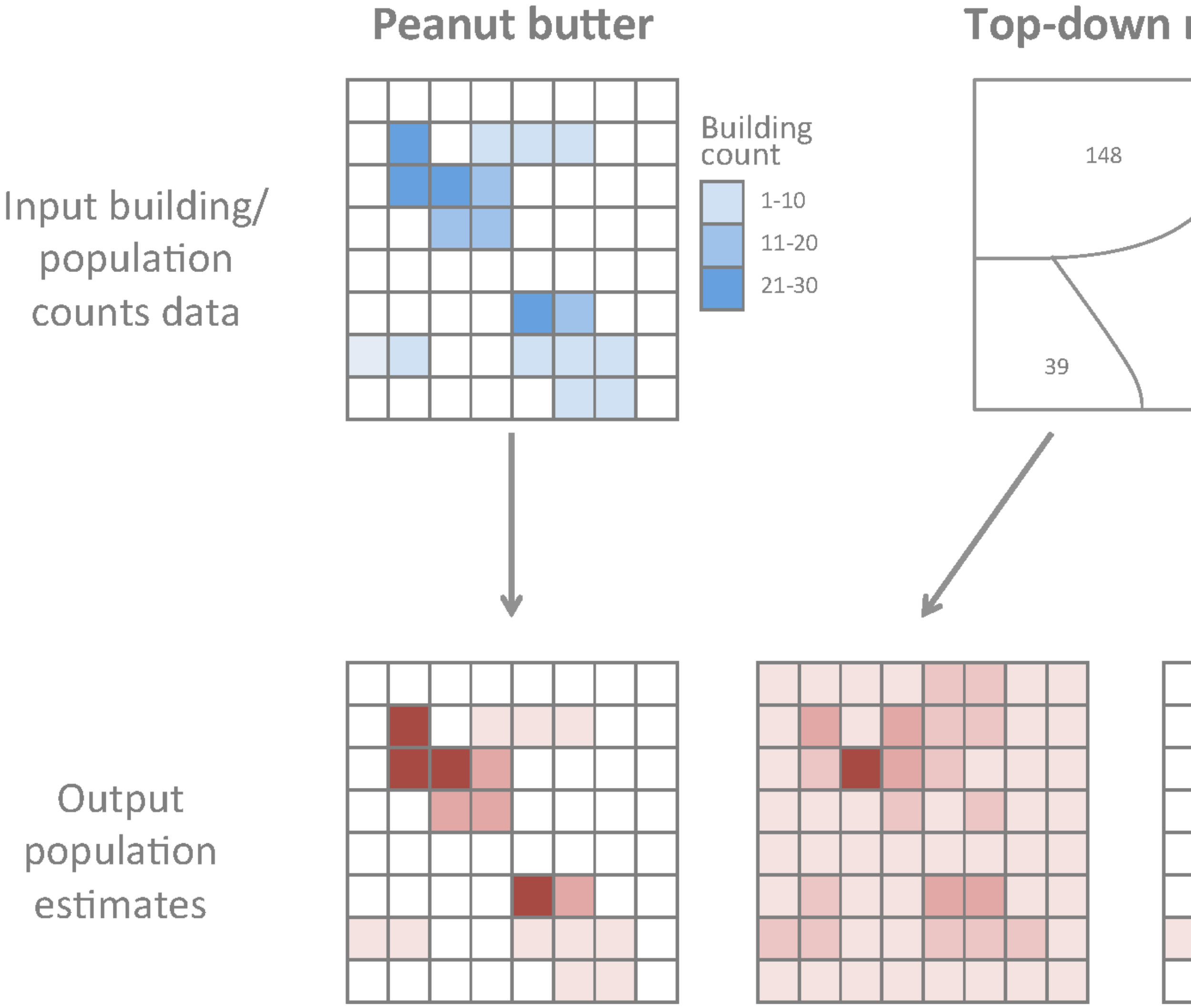
Peanut butter
Top-down r
Input building/ population counts data
Building count
1-10
11-20
21-30
148
39
Output population estimates

Figure 1.1: Comparison of inputs and outputs for four types of gridded population estimates (from bottom-left to bottom-right): peanut butter, top-down, top-down constrained, and bottom-up.

See the Bottom-up Models section for more details.

1.3 Peanut Butter

The peanut butter method spreads user-provided estimates of people per building evenly (like peanut butter) among buildings. This is a quick and simple approach that utilizes high resolution maps of building footprints, assuming that the same number of people live in each building across entire regions or settlement types (e.g. urban and rural). WorldPop partners with Maxar Technologies, Ecopia.AI, and the Bill and Melinda Gates Foundation for recent maps of buildings (**?**, **?**).

This is a good approach when rapid-response population estimates are needed and no suitable data are available for more data-driven methods. The peanut butter method ignores spatial variation of people per building within each region or settlement type and there is often no objective basis for assessing uncertainty.

The peanut butter method can be applied in aggregation mode or disaggregation mode:

- **Aggregation mode** relies on expert opinion to define the average number of people per building footprint in each region or settlement type.
- **Disaggregation mode** calculates the average number of people per building footprint based on user-provided population totals in each region.

In both cases, the user-provided estimates of people per building are mapped to every building footprint. Population totals within each 100 m grid cell or administrative unit are calculated by summing across all of the relevant building footprints.

Try the peanut butter method yourself using the peanutButter web application or the peanutButter R package (**?**).

1.4 Comparison of Data

Within these three broad categories, WorldPop develops a variety of methods for producing gridded population estimates with different characteristics. Selecting the right population estimation method depends on project requirements (e.g. time and resource availability) and data availability. Table **??** provides a checklist of features for different types of WorldPop gridded population estimates.

Comparison of data characteristics for different types of gridded population estimates. Columns list types of population estimates and rows provide their characteristics. X marks a feature of a data set and ? means that a feature is optional. Variations of the top-down method: 1) Census Projections: basic top-down based on projected census totals, 2) UN-adjusted: includes a post-hoc adjustment to match UN population estimates for administrative units, 3) Constrained: population estimates are constrained to grid cells where buildings occur (UN-adjustment optional), and 4) Custom: WorldPop researchers developed a custom top-down model using bespoke data for an individual country.

Top-down

Peanut Butter

Census Projections

UN-adjusted

Constrained

Custom

Bottom-up Bayesian

Input Population Data

Expert opinion (e.g. people per building)

X

Population totals for admin units

X

X

X

X

Geolocated household survey data

X

Output Data

Gridded population estimates (~100 m)

X

X

X

X

X

X

Sum to match projected census totals

?

X

X

?

Adjusted to match UN admin unit totals

?

X

?

Constrained to buildings

X

X

X

X

Include estimates of uncertainty

X

Access

www.worldpop.org

X

X

X

ftp.worldpop.org

X

X

X

X

X

wopr.worldpop.org

X

X

apps.worldpop.org/woprVision

X

X

apps.worldpop.org/peanutButter

X

With any of these methods, gridded estimates of total population can be divided into specific age-sex groups using WorldPop's pre-existing gridded age-sex proportions. With bottom-up methods, it is also possible to estimate age-sex proportions directly from household survey data when it is available. See section Age-sex Mapping for more information.

Contributing

This chapter was written by Doug Leasure, Claire Dooley, *[contributors, please add your name here]*. Funding for the work described in this chapter was provided by *[please add funders and grant numbers here]*.

Suggested Citation

WorldPop. 2020. Introduction: Gridded Population Estimates. In *WorldPop Book of Methods, Vol. I: Gridded Population Estimates*. WorldPop, University of Southampton. 26 November 2020, https://docs.worldpop.org

License

Chapter 2

Data Access

> **Note:** This chapter is currently being drafted. To make suggestions, please raise an issue in the bookworm repository. To make direct edits to the source code, please submit a pull request to merge your code with the "dev" branch. Include the tag @doug-leasure in your issues or pull requests to notify Doug.

You can access WorldPop data sets in a variety of ways that could include downloading individual files from WorldPop.org, downloading in bulk from the WorldPop FTP server, or creating dynamic links to population data from your own web server using REST API. The best way to access population estimates may depend on how you intend to use the data and the characteristics of the specific data set that you are accessing.

2.1 Websites

WorldPop.org is the central location to access WorldPop data that has been produced across a range of projects. This includes gridded population estimates for most countries from the WorldPop Global Project (WorldPop et al **?**) along with gridded estimates of births, pregnancies, age-sex structure, urban change, development indicators and other population-related variables.

We are also now developing the WorldPop Open Population Repository to publish bespoke population data for individual countries, to provide Bayesian estimates of uncertainty, and to link these data sets to web applications and other tools.

2.2 Web Applications

WorldPop web applications are available from apps.worldpop.org.

These applications allow you to explore population data using interactive web maps and other tools to maximize the information you get from population data.

Global Demographics Portal

The WorldPop Demographics app is available at portal.worldpop.org/demographics. This application allows you to visualize age-sex proportions estimated for small areas and mapped across every country.

Global Population Data Portal

The WorldPop Global Data Portal is available from portal.worldpop.org. This web portal allows you to visualize and download top-down gridded population estimates for most countries in the world (WorldPop et al **?**).

peanutButter

The peanutButter application (**?**) is available from apps.worldpop.org/peanutButter. This application allows you to produce gridded population estimates from building footprints using the peanut butter method. This simple approach requires you to provide estimates of the average number of people per building in each settlement type (e.g. urban and rural). Your estimates are mapped across buildings using high resolution maps of building footprints (**?**) that are based on recent satellite imagery.

woprVision

The woprVision application (**?**) is available from apps.worldpop.org/woprVision. This app is an interactive web map that allows you to query population estimates for specific locations and demographic groups from the WorldPop Open Population Repository. This can be used to download population data, query population estimates for specific locations and demographic groups, and retrieve probabilistic Bayesian estimates of uncertainty.

2.3 FTP Server

The WorldPop FTP server provides a good resource for downloading files in bulk. Most data available from the “DATA” tab at worldpop.org can also be downloaded from the “GIS” folder on the FTP server.

This includes gridded population estimates produced using the top-down method for most countries in the world as well as the gridded spatial covariates used for modelling (WorldPop et al. **?**).

The “repo” folder on the FTP server contains permanent archives of data sets and code from worldpop.org sub-domains including “wopr”, “apps”, and “docs”. For example, the “wopr” sub-directory contains archived data from wopr.worldpop.org.

2.4 GIS Plugins

wpgpDataAPD
This Esri plugin / ArcPy Python toolbox allows you to download WorldPop gridded population estimates produced using the top-down method for most countries globally (WorldPop et al **?**) directly from Esri ArcGIS software. See wpgpDataAPD on GitHub.

wpgpDataQPD
This QGIS plugin allows you to download WorldPop gridded population estimates produced using the top-down method for most countries globally (WorldPop et al **?**) directly from QGIS software. See wpgpDataQPD on GitHub.

2.5 R Packages

peanutButter
This package allows you to create your own gridded population estimates using the peanut butter method and high resolution building footprints (**?**). See peanutButter on GitHub.

wopr
The wopr package (**?**) allows you to download bottom-up gridded population estimates from the WorldPop Open Population Repository from your R console and submit spatial queries (i.e. points or polygons) to retrieve population estimates for specific locations and demographic groups with statistical estimates of uncertainty. It also allows you to run the woprVision web application from your R console. See wopr on GitHub.

wpgpCovariates
This package provides access to gridded spatial covariates (WorldPop et al. **?**) for most countries. See wpgpCovariates on GitHub.

wpgpDownloadR
This package provides access to top-down gridded population estimates (WorldPop et al. **?**) for most countries from your R console. See wpgpDownloadR on GitHub.

2.6 Python Packages

wpgpDownloadPy
This Python package provides access to top-down gridded population estimates (WorldPop et al. **?**) from the Python console. See wpgpDownloadPy on GitHub.

2.7 REST API

REST API is a way for computers to communicate with one another to request data downloads or query databases. Many WorldPop datasets can be accessed using REST API requests. This makes it possible to automatically sync remote servers with WorldPop population data and to develop web applications that use API to query WorldPop servers.

WOPR API
This can be used to query bottom-up population estimates from the WorldPop Open Population Repository. These API endpoints can be used to download entire data sets for each country or to submit spatial queries to the WorldPop server to request population estimates for specific locations and demographic groups. The WOPR API endpoints return Bayesian estimates of uncertainty for all population estimates. See the chapter WOPR API for more information.

WorldPop API
This can be used to download top-down gridded population estimates from the WorldPop Global Project (WorldPop et al **?**). See WorldPop API documentation for more information.

Contributing

This chapter was written by Doug Leasure, *[contributors, please add your name here]*. Funding for the work described in this chapter was provided by *[please add funders and grant numbers here]*.

Suggested Citation

WorldPop. 2020. Introduction: Data Access. In *WorldPop Book of Methods, Vol. I: Gridded Population Estimates*. WorldPop, University of Southampton. 26 November 2020, https://docs.worldpop.org/bookworm

Population Mapping Methods

This section provides methodological details for the various approaches that WorldPop has developed for producing gridded population estimates.

Chapter 3

Top-down Models

Note: This chapter is currently being drafted. To make suggestions, please raise an issue in the bookworm repository. To make direct edits to the source code, please submit a pull request to merge your code with the "dev" branch. Include the tag @bondarenkom in your issues or pull requests to notify Max and/or @asoriche to notify Ale.

3.1 CIESIN projections

3.1.1 Global

3.1.2 Individual countries

3.2 UN-adjusted

3.3 Constrained to building footprints

3.4 Conclusion

Contributing

This chapter was written by *[contributors, please add your name here]*. Funding for the work described in this chapter was provided by *[please add funders and grant numbers here]*.

Suggested Citation

Doe J, … . 2020. Population Mapping Methods: Top-down Models. In *WorldPop Book of Methods, Vol. I: Gridded Population Estimates.* WorldPop, University of Southampton. 26 November 2020. https://docs.worldpop.org

Chapter 4

Bottom-up Models

> **Note:** This chapter is currently being drafted. To make suggestions, please raise an issue in the bookworm repository. To make direct edits to the source code, please submit a pull request to merge your code with the "dev" branch. Include the tag @doug-leasure in your issues or pull requests to notify Doug.

Bottom-up population modelling methods (**??**) use geolocated household survey data from a sample of locations to fit statistical models that estimate population sizes for unsampled areas based on relationships with spatial covariates. WorldPop develops customized statistical models for individual countries to make the best use of available survey data and to provide robust estimates of uncertainty.

This is a good approach when there has not been a recent or complete national census but there are recent geolocated household survey data available. This approach provides Bayesian estimates of uncertainty but requires more detailed input data and more time to develop the models.

4.1 Input Data

Bottom-up methods require a few key types of input data:

1. Population data

2. Settlement map

3. Geospatial covariates

4. Administrative boundaries

4.1.1 Population Data

Population data for bottom-up methods generally must include counts of people in clearly defined georeferenced areas. A polygon shapefile with the boundary of each enumeration area and the total population within each area is ideal. There are a few potential sources for these data:

- Partial census results
- Microcensus surveys designed for population modelling (a random sample of locations where enumeration is carried out)
- Pre-survey listing data from routine household surveys (e.g. DHS, LSMS, MICS)

Point locations of buildings and/or households within enumeration areas are sometimes collected during census and survey field work. These data can be very useful because they provide higher resolution information about population patterns, but they are not required. Pre-survey listing data can be very useful, especially if surveys were recently conducted in areas that were inaccessible to census enumerators. If pre-survey listing data from household surveys are used, additional information about the site selection will also be required. If the household survey used a sampling design in which survey locations were selected with probabilities proportional to population size (PPS), then it will be necessary to obtain the weights used for PPS sample design.

4.1.2 Settlement Map

A settlement map identifies areas where residential structures occur. It may also classify areas into settlement types such as urban, peri-urban, rural, slums, commercial, industrial, etc (see Settlement Classification). This information may be in the form of:

- Building locations (points)
- Building footprints (polygons)
- Gridded map identifying pixels that contain buildings (raster)

These data could be derived from several sources:

- Satellite imagery

- Pre-census cartography

- Building points and footprints can be purchased commercially
- Gridded derivatives of building footprints are freely available for some countries (?)

If there is no classification of settlement types available, building points or building footprints could be directly used to identify different settlement types based on the patterns of building locations (??). There are also freely available global settlement maps, but quality from global data sets varies strongly among countries, with the smallest settlements often missing, so this would need to be considered before committing to any publicly available global settlement map.

Additional data about each building can be very beneficial for population modelling such as building area, height, or use (i.e. residential, commercial, mixed). Classifying individual buildings as residential or non-residential (??) can sometimes be accomplished with existing public data from Open Street Maps and other sources. While these additional data would improve population estimates, they are not required.

4.1.3 Geospatial Covariates

Geospatial covariates are spatial data (e.g. GIS data) with national coverage that describe any variable that may be correlated with population densities. There are many suitable datasets that are publicly available, including some produced by WorldPop for this purpose (??).

For example, a digital map of road networks (a line shapefile) could be used to calculate road densities which may correlate with population densities. Or, global satellite-derived nighttime lights data sets (raster files) may correlate with population densities in some areas. Administrative records could also be useful such as electricity usage for each administrative unit (polygon shapefile). Locations of public facilities such as schools (a point shapefile) can also be very informative. If the number of students attending each school is known, that would also likely add to the accuracy of population estimates.

There are an almost infinite number of possible geospatial covariates. Many of them are publicly available, so identifying these data sets is not necessarily required to initiate population modelling. But, identifying good quality covariates (i.e. those that are strongly correlated to population density) that are comprehensive with national coverage can significantly improve the accuracy of population estimates.

4.1.4 Administrative Boundaries

Administrative boundaries could include regions, states (provinces), and/or local government areas. These administrative units are often nested within one another. Administrative units can be used by the model as a covariate to improve estimates of population densities. Administrative units can also be used to summarize model results, providing population totals for each administrative unit.

4.2 Statistical Models

WorldPop develops customized Bayesian models to make the best use of available data for specific countries and to accurately quantify uncertainty associated with the population estimates.

Bayesian models generate population estimates as probability distributions known as "posteriors". You can see examples of posterior probability distributions for population estimates in the woprVision web application. We use the mean value of the posterior probability distribution as the expected value for the population estimate. Variance around the mean represents uncertainty in the population estimate.

Uncertainty in population estimates may be caused by several factors. Sometimes uncertainty results from sampling error associated with small sample sizes (i.e. not many household survey clusters in the area). Uncertainty may also represent true variation in population densities from neighborhood to neighborhood that simply could not be explained by the covariates in the model. Uncertainty may also be related to the structure of the statistical model itself. To reduce uncertainty in a model, you must weigh the cost-benefits for: A) collecting more household survey data, B) finding better covariates to predict population densities, or C) revising the model structure. Revising the model structure is by-far the easiest and this is one reason why the flexibility of Bayesian models is so important.

4.2.1 Software

A quick note about software before we get into the models themselves. The R programming language for statistical computing (**?**) is ideal for fitting Bayesian models. There are a number of software packages available, but a few that we regularly use are:

1. STAN software (**?**) with the rstan R package (**?**)

2. JAGS software (**?**) with the runjags R package (**?**)

3. INLA R package (**?**)

If you are new to Bayesian modelling, we recommend starting with STAN because it provides full flexibility to customize your models, it has excellent documentation (mc-stan.org) and it is computationally more efficient than JAGS. If you are already familiar with the BUGS or JAGS languages, you can build all of the models described below using either software. If you want to build geostatistical models (see Geostatistical Models), INLA (R-INLA.org) is the preferred software because it is more computationally efficient than JAGS or STAN for estimating high-dimensional spatial covariance parameters, although it is less flexible for building customized hierarchical models.

4.2.2 Simple Model to Start

A simple linear regression can be written as:

$$y_i \sim Normal(\mu_i, \sigma)\mu_i = \alpha + \beta x_i$$

where y_i is the value of the response variable at location i and x_i is the predictor variable (a.k.a. covariate). These two variables represent observed data and all of the other variables represent model parameters that we will estimate using STAN, JAGS, or INLA (software described above).

μ_i is the expected value of the response variable based on the covariate value x_i at a given location. It is the mean of the normal distribution. Random noise that could result in the observed value being different than the expected value (i.e. residual variance, uncertainty) is represented by σ (i.e. standard deviation). The regression coeffecient β (i.e. regression slope) estimates the effect of the covariate on the expected value, and α (i.e. regression intercept) is the expected value for the response variable when the covariate is equal to zero.

The first line of the model is the stochastic model (i.e. it includes random noise) and the second line is deterministic (i.e. it always generates the same output for a given input). The selection of a normal distribution (a.k.a. Gaussian) in the stochastic portion of the model should be based on characteristics of the response variable. A normal distribution represents continuous numbers that can be negative or positive.

For population modelling, our response variables are counts of people N_i that are always positive integers, so we need to choose a more appropriate stochastic model. We can modify the Gaussian linear regression above into a Poisson regression:

$$N_i \sim Poisson(\mu_i)log(\mu_i) = \alpha + \beta x_i \tag{4.1}$$

This is a generalized linear model with a log-link function (**?**). The log-link function ensures that μ_i is always positive, and the Poisson distribution produces positive integers. Now we have an appropriate deterministic regression and stochastic model for population counts.

4.2.3 Bayesian Priors

To implement this model Eq. (**??**) in a Bayesian context, we need to define priors for α and β. Priors are probability distributions that represent our prior knowledge about the range of possible values for parameters in the model. Priors must be specified for any "root node" parameters, those that do not show up on the left side of any probability statements in the model. Probability distributions used as priors are usually very disperse flat priors so that they do not influence the posterior parameter estimates. In general, we try to specify priors that are informative enough to define a realistic range of possible values for the parameter but vague enough to allow the observed data to have dominating influence on the parameter estimates.

For the model in Eq. (**??**), we could choose uninformative flat priors:

$$\alpha \sim Uniform(-10, 10)\beta \sim Uniform(-10, 10)$$

On the log-scale, this is a range from near zero to over 22,000.

Or, we may prefer more informative priors:

$$\alpha \sim Normal(0, 5)\beta \sim Normal(0, 1)$$

The relative influence of priors depends on the scale of the response variable and the structure of the model. It is good practice to test the relative influence of various priors on the posterior parameter estimates before making a decision.

In this chapter, we will assume that you understand Bayesian priors and we will not explicitly specify the priors in our examples unless the prior selection is noteworthy.

4.2.4 Hierarchical Core Model

A hierarchical model is one where the output (left side of equation) from one stochastic model serves as the input (right side) of another. Building on the Poisson regression in Eq. (**??**), we can make a hierarchical model that incorporates population density D_i:

$$N_i \sim Poisson(D_i A_i) D_i \sim LogNormal(\bar{D}_i, \sigma)\bar{D}_i = \alpha + \sum_{k=1}^{K} \beta_k x_{i,k} \tag{4.2}$$

where A_i is observed data measuring total settled area within location i. If area is measured in hectares, then D_i would be people per hectare. $\bar{D}_i$ is the expected population density on the log scale (i.e. the mean of the log-normal distribution), and σ is the residual variance term. K is the total number of covariates included in the model.

This hierarchical formulation has several advantages over the simple Poisson regression from Eq. (**??**):

1. It adds a residual variance term σ that allows for over-dispersion of the Poisson,
2. Covariates are now predicting population density rather than counts, and
3. The log-normal replaces the log-link function (acting as a stochastic log-link).

Over-dispersion means that the model can now accommodate more residual variance in population counts than could be modelled with a Poisson distribution alone (because Poisson doesn't have a variance parameter). In addition, making the covariates predictors of population density rather than population counts avoids the confounding effect of area. For example, two locations with identical covariate values and population densities could have very different population counts if the total amount of settled area is different. The hierarchical model explicitly accounts for this multi-level process.

Eq. (**??**) will serve as the core likelihood model for many of the model customizations described below.

4.2.5 Age-sex Structure

We can incorporate an age-structured sub-model if the household survey data contain counts $M_{i,g}$ of people in each age-sex group g at each location i. These data allow us to estimate a population pyramid (i.e. proportions of the population in each age-sex group) and to produce age-sex-specific population estimates. A multinomial model can be added to Eq. (**??**) to achieve this:

$$M_{i,g} \sim Multinomial(\theta_{r,g}, N_i)\theta_{r,g} \sim Dirichlet(rep(1,g)) \tag{4.3}$$

where N_i is the total population at location i from Eq. (**??**). The population pyramid $\theta_{r,g}$ is being estimated independently for each region r with a flat Dirichlet prior. The Dirichlet prior enforces the assumptions that individual elements of $\theta_{r,g=1:G}$ are between zero and one and that they sum to one across all age-sex groups g.

4.2.6 Random Intercept

Random effects are regression coefficients (e.g. α and β above) that are dependent on other parameters. All of the regression coefficients shown above were fixed effects because they were not dependent on other parameters. Models that contain random effects are sometimes called mixed effects models because they contain fixed and random effects. Mixed effects models may have random intercepts, random slopes, or both.

An example of a random intercept in a population model could be a regression intercept α (i.e. average population density) that is estimated separately for urban and rural areas in a way that accounts for the correlation between the two. We can adjust Eq. (**??**) to have this random intercept α_t:

$$N_i \sim Poisson(D_i A_i) D_i \sim LogNormal(\bar{D}_i, \sigma) \bar{D}_i = \alpha_t + \sum_{k=1}^{K} \beta_k x_{i,k} \alpha_t \sim Normal(\eta, \theta) \tag{4.4}$$

where t is the settlement type (i.e. urban or rural) that location i belongs to. η and θ are the mean and standard deviation of α among settlement types. The correlation between α for the two settlement types is explicitly modelled because they are drawn from the same distribution, but these parameter estimates will still differ based on their fit to the data from each settlement type. This is a random intercept by settlement type and it can help to account stratified sampling that household surveys often use to collect population data.

We can extend this concept to a random intercept by settlement type t and region r to account for additional spatial correlation where population densities from the same region are more similar to one another than population densities from different regions. Regions r could be defined as states or local government areas. This two-level random intercept $\alpha_{t,r}$ (by settlement type and region) could be included as:

$$N_i \sim Poisson(D_i A_i) D_i \sim LogNormal(\bar{D}_i, \sigma) \bar{D}_i = \alpha_{t,r} + \sum_{k=1}^{K} \beta_k x_{i,k} \alpha_{t,r} \sim Normal(\breve{\alpha}_t, \theta_t) \breve{\alpha}_t \sim Normal(\bar{\alpha} \tag{4.5}$$

where $\breve{\alpha}_t$ and θ_t are the mean and standard deviation (for each settlement type) of regression intercepts $\alpha_{t,r}$ among regions. At the national level, $\bar{\alpha}$ and η are the mean and standard deviation for $\breve{\alpha}_t$.

This hierarchical random intercept can help to account for:
- Sampling that is stratified by settlement type, and - Spatial autocorrelation within regions.

4.2.7 Hierarchical Variance

Similar to the hierarchical random intercept above, we can also use hierarchical variance by settlement type and region. This allows us to map uncertainty to see where residual variance is the greatest and giving more realistic ranges of uncertainty around population estimates in different regions and settlement types. We could modify Eq. (**??**) to have hierarchical variance $\sigma_{t,r}$:

$$N_i \sim Poisson(D_i A_i) D_i \sim LogNormal(\bar{D}_i, \sigma_{t,r}) \bar{D}_i = \alpha + \sum_{k=1}^{K} \beta_k x_{i,k} \sigma_{t,r} \sim HalfNormal(\breve{\sigma}_t, \theta_t) \breve{\sigma}_t \sim HalfNormal(\bar{\sigma}, \tag{4.6}$$

Half-Cauchy distributions are also often recommended for modelling hierarchical variances rather than the Half-Normal that we have shown here (**?**). Hierarchical variances can lead to convergence issues and care must be taken to specify priors that result in good convergence without being too influential on the posterior parameter estimates. It is often necessary to simplify the variance structure (e.g. fewer settlement types, or regions, or dropping one level entirely), especially if the sample size is low in some regions and/or settlement types.

4.2.8 Weighted-likelihood

Household surveys often implement a weighted sampling design known as PPS, or Probability Proportional to Size. This means that the probability of a location being selected for the survey is not random, it is dependent on the number of people (or households) in that area. Household surveys use weighted sampling to achieve a representative sample of households. If they used spatial random sampling, the results would be biased towards rural areas because urban areas occupy less space on the landscape.

To use these data for population modelling, it is necessary to account for the bias that weighted sampling can introduce to avoid overestimating average population densities. Suppose sample weights w_i were used to collect a weighted sample of locations from a national sampling frame. We can build a weighted-likelihood model that incorporates these weights to provide unbiased estimates of population densities. The first step is to calculate inverse weights and scale them to sum to one:

$$m_i = \frac{w_i^{-1}}{\sum_{i=1}^{1} w_i^{-1}} \tag{4.7}$$

The scaled inverse weights m_i (or “model weights”) are then used to weight individual samples in the likelihood by adjusting the variance term σ_i:

$$N_i \sim Poisson(D_i A_i) D_i \sim LogNormal(\bar{D}_i, \sigma_i) \sigma_i = \sqrt{\frac{1}{m_i \theta^{-2}}} \tag{4.8}$$

where θ is an estimated parameter that is a component part of the variance, together with the model weights m_i. Notice that the standard deviation for the log-normal σ_i is now location-specific. This results in unbiased estimates of the mean and variance because it gives more weight in the likelihood to locations that had lower probabilities of being included in the sample (e.g. for PPS household survey designs this would be locations with fewer people). The regression model for $\bar{D}_i$ is not shown but it could be setup like Eq. (**??**).

The sample weights w_i are often not known for unsampled areas where population predictions are needed. Because of this, we need to derive a weighted average value for the variance term that is not location-specific:

$$\bar{\sigma} = \frac{\sum_{i=1}^{I} \sigma_i \sqrt{m_i}}{\sum_{i=1}^{I} \sqrt{m_i}}$$

This is a weighted average of σ_i across locations i (essentially factoring out the model weights m_i). Model predictions of population density $\hat{D}_i$ in locations where sampling weights w_i are unknown would be produced from:

$$\hat{D}_i \sim LogNormal(\bar{D}_i, \bar{\sigma})$$

4.2.9 Geostatistical Models

Geostatistics is a form of spatial statistics that explicitly model a continuous spatial phenomenon when observations are accurately georeferenced at particular sites (such as from a GPS location in a survey). Geostatistical models can help to estimate the outcome in unobserved locations, with the expectation that nearby locations are more similar than distant locations. While geostatistical modelling includes interpolation or smoothing methods such as Kriging, a model-based geostatistical approach (**?**) makes it possible to incorporate spatial position into a statistical framework similar to Eq. (**??**). The spatial information from the observations' locations (in addition to observed covariate data) can improve the accuracy of population estimates. WorldPop has utilised geostatistical modelling approaches to produce population estimates for Afghanistan (CITATION), to map the proportion of the population under 5 years of age (**?**), produce high-resolution poverty estimates (**?**), and to estimate vaccination coverage (**?**).

The general form of the model-based geostatistics framework is a mixed-effects regression model. This model includes fixed covariate effects plus a spatially correlated random effect for modelling spatial variation, to give