KU LEUVEN

FACULTY OF SCIENCE

# Refining Population Mapping with Nighttime Lights:

## A Bayesian Spatiotemporal Approach with SPDE-INLA

**Yueyang YI (r0730064)**

Supervisor: prof. dr. Thomas Neyens
Katholieke Universiteit Leuven

Master's thesis submitted in fulfillment
of the requirements for the degree in
Master of Science in Statistics and Data Science

Academic years 2020-2021-2022-2023

# Preface

First and Foremost, I would like to express my sincerest gratitude to my supervisor, prof. dr. Thomas Neyens, whose insights and knowledge guide me throughout the journey. Especially, I would like to thank him for his enthusiasm, patience and professionalism. His warm, persistent and unwavering supports make this endeavour become possible.

Besides, I would like to thank my classmates: Yabiao Peng and Simin Xie, for helping me adapt to study in Leuven; Ziyue Zhu and Changsheng Chen, for their tremendous encouragements; Chunbo Wang and Xinyi Li, for their timely assistance with my travel to Leuven; Qiming Sun and Runyang Wang, for the happy days that we spent together.

Finally, I would like to thank my parents for their unconditional love and supports since the beginning of my life. The journey here has been long, but I will keep calm and carry on.

# Summary

This master's thesis aims to develop an approach that projects existing gridded population estimates based on "top-down" methods and available for free to any target pair of spatial and temporal resolutions, with assistance from easy-to-measure or hard-to-change geospatial ancillary data. If the projected "top-down" population estimates based on the existing gridded population data products available for free hold the same or even better quality than the ones achieved in a "bottom-up" approach based on the limited small-scale "microcensus" household surveys, a "free lunch" can be enjoyed. That means, under such a circumstance, the user does not have to pay for the "microcensus", as paying extra does not bring about better predictive performances. This idea is the first attempt trying to assess the border between "top-down" and "bottom-up" methods, so as to find the minimum sample size that lets the predictive performances of "bottom-up" estimates be better than the projected large-scale population data products available for free. It is also the first time to use ancillary data available for free and "top-down" global population data products available also for free for calculating prior information for "bottom-up" estimates. The proposed approach represents a significant step towards a combination of mainstream "top-down" and rising "bottom-up" approaches that makes the most of census and survey data.

# Acronyms

**AI** Artificial Intelligence.

**BAG** Basic Registration of Addresses and Buildings.

**BRP** Basic Registration of Persons.

**CBS** Centraal Bureau voor de Statistiek.

**DEM** Digital Elevation Models.

**DMSP** Defense Meteorological Satellite Program.

**DN** Digital Number.

**DNB** Day Night Band.

**FEM** Finite Element Method.

**GBR** Geographical Basic Register.

**GF** Gaussian Field.

**GHS-BUILT-S** GHS Built-up Surface Grid.

**GHS-POP** Global Human Settlement Population Grid.

**GHS-SMOD** GHS Settlement Model Layers.

**GHSL** Global Human Settlement Layer.

**GMRF** Gaussian Markov Random Field.

**GPW** Gridded Population of the World.

**GRF** Gaussian Random Field.

**HPD** Highest Posterior Density.

**IIVS** Integrated Income and Asset Statistics.

**INLA** Integrated Nested Laplace Approximation.

**JAGS** Just Another Gibbs Sampler.

**LGCP** Log-Gaussian Cox Process.

**MCMC** Markov Chain Monte Carlo.

**MERIT DEM** Multi-Error-Removed Improved-Terrain DEM.

**ML** Machine Learning.

**NTL** Nighttime Lights.

**OLS** Operational Linescan System.

**PC** Penalised Complexity.

**PCC** Pearson Correlation Coefficient.

**PIT** Probability Integral Transform.

**RMSE** Root Mean Square Error.

**SD** Standard Deviations.

**SPDE** Stochastic Partial Differential Equations.

**VIIRS** Visible Infrared Imaging Radiometer Suite.

**WorldPop** WorldPop Global High Resolution Population Denominators.

**WOZ** Valuation of Immovable Property.

# List of Figures

# List of Tables

# Contents

# Chapter 1

# Introduction

This master's thesis consists of seven chapters: Chapter 1 introduces "top-down" population mapping, "bottom-up" population mapping and the research purpose of this master's thesis; Chapter 2 introduces the data used for implementing the proposed approach and the idea of simplifying the use of ancillary data; Chapter 3 introduces the statistical basis; Chapter 4 and Chapter 5 show the results from the tests of the idea of projecting existing gridded population estimates to any spatial resolution, the idea of combining "top-down" and "bottom-up" approaches and the idea of assessing the border between "top-down" and "bottom-up" methods, with the projected "top-down" gridded population data considered to be precise and less precise estimates on the real population data respectively; Chapter 6 shows the results from the test of the idea of projecting existing gridded population estimates to any temporal resolution; Chapter 7 concludes the current works and recommends the further works.

## 1.1 "Top-Down" Population Mapping

Accurate regional population mapping at local levels is important for policymakers to practice the approaches developed by research in many disciplines of public interest, for instance, measuring urban economic development (e.g., Zhao et al., 2017; Henderson et al., 2021), assessing large-scale disaster risk (e.g., Winsemius et al., 2013; Ward et al., 2013) and analysing the population-level public health issues (e.g., Giani et al., 2020; Carleton et al., 2021). A population mapping approach that allocates aggregated population counts (e.g., census data) to fine grid squares (e.g., at a 1 km spatial resolution) via some appropriate methods (e.g., dasymetric, statistical and Machine Learning (ML)-based methods) is called a "top-down" approach. Large-scale gridded population data products produced with "top-down" methods are often selected as primary input data (e.g., Ward et al., 2020; Henderson et al., 2021), as they appropriately allocate population numbers at some upper level of spatial scale (i.e., coarse administrative units) to fine grid squares that cover sufficiently large geographic extents including data-scarce developing countries.

Leyk et al. (2019) related the data products' fitness for use to the target applications in their thorough review. Table 1.1 summarises the availability of spatial and temporal resolutions of several large-scale population data products. The user would often find there is a mismatch between the operational scale on which a certain process or phenomenon of interest is modelled and the analytical scale on which the grid cell is defined by a certain spatial resolution (e.g., Maclaurin et al., 2015; MacManus et al., 2021). Temporal mismatches between pop-

ulation and any of the intrinsically time-varying ancillary data inputs may also be critical for the intended application which requires a high degree of temporal currency (Sun et al., 2017; Leyk et al., 2019; Archila Bustos et al., 2020). It prompts the use of temporally implicit or invariant ancillary data in the modelling process (Gaughan et al., 2016), but projecting forward or backward from census data to a target time of interest still remains a challenge (Leyk et al., 2019). The gridded population estimates have thus to be understood as an approximation over a period of time.

Table 1.1: Availability of spatial and temporal resolutions of several existing large-scale population data products

| Database | Temporal resolutions | Spatial resolutions |
|---|---|---|
| GPW v4.11 | 2000; 2005; 2010; 2015; 2020 | 30 arc-seconds |
| GRUMP v1 | 1990; 1995; 2000 | 30 arc-seconds |
| LandScan | annual; 2000-2016 | 30 arc-seconds |
| WorldPop | annual; 2000-2020 | 3 arc-seconds; 30 arc-seconds |
| GHS-POP | 1975; 1990; 2000; 2015 | 3 arc-seconds; 30 arc-seconds |

Whenever a population data product does not provide a specific pair of spatial and temporal resolutions required by a specific application, a naïve solution is to switch to the other population data products that have this specific pair. Whenever a population data product provides a finer resolution than the ones provided by the other population data products, a naïve idea is to choose the population data product that holds the finer resolution. However, a population data product cannot be easily considered as a substitute for the other ones, because (i) a finer spatial resolution does not necessarily imply greater accuracy (MacManus et al., 2021), as uncertainty may be associated with the original census, the areal aggregation of the inputs and the statistical models (Nagle et al., 2014; Doxsey-Whitfield et al., 2015; Sinha et al., 2019); (ii) currently, neither standardised measures nor the concept of fitness for use is available to determine whether or not a data product meets the need of an intended use due to the complexity of uncertainty (Leyk et al., 2019), no matter how comprehensive the modelling techniques that data product employs are (Archila Bustos et al., 2020); (iii) what the population modelled represents may also be fairly different, e.g., different population concepts or population groups in different regions or countries defined by different census inputs (see Leyk et al., 2019; Archila Bustos et al., 2020; MacManus et al., 2021); for instance, using de jure/de facto (or legal/present/nighttime) population estimates over ambient (or daytime) ones would possibly lead to dramatically different results. There are only several large-scale population data products available for free, while each of them is produced with different methods for different population representations. Hence, each of them is considered to be unique, valuable and irreplaceable. Table 1.2 summarises "top-down" methods and population representations of several large-scale population data products.

A second naïve idea is to reproduce those large-scale population data products at the resolution required, on the basis of the same "top-down" methods. However, those large-scale population data products are considered to be irreproducible, because (i) the "top-down" methods often require a variety of ancillary variables for population redistribution, and it is hard to collect such data that can be used in a large-scale population grid production consistently in terms of both space and time (Leyk et al., 2019); (ii) the user of the large-scale population

Table 1.2: "Top-down" methods and population representations of several existing large-scale population data products

| Database | Modelling methods | Representations |
|---|---|---|
| GPW v4.11 | areal weighting | de jure/de facto |
| GRUMP v1 | dasymetric | de jure/de facto |
| LandScan | smart interpolation | ambient (daytime) |
| WorldPop | statistical/dasymetric | de jure/de facto |
| GHS-POP | dasymetric refinement with built-up info | de jure/de facto |

data products produced with ML-based methods (e.g., WorldPop data) may find the range of values in the training observations does not allow extrapolation beyond the bounds of the population densities observed in the original training data (Stevens et al., 2015); the user who tries to reproduce them at the resolution required has thus to provide a large amount of new training data at the resolution required; (iii) the algorithms used by some population data products are proprietary (e.g., LandScan data).

Thus, projecting the large-scale "top-down" population estimates well filed as the data products available for free to the required resolution with a limited amount of ancillary data that can be easily obtained would be a solution to dealing with spatial and temporal mismatches.

## 1.2 "Bottom-up" Population Mapping

Instead of distributing the best available census data, recently emerging "bottom-up" approaches estimate up-to-date gridded population in countries without a complete and recent census, based on limited household surveys in randomly selected small defined areas (also known as "microcensus") and several geospatial covariates available nationwide. Such "microcensus" data can be collected relatively rapidly and at a fraction of the cost of a complete national census. Geospatial covariates are normally required to be strongly correlated to population distribution and available consistently across the target areas. Wardrop et al. (2018) summarised the challenges to "top-down" disaggregation and the feasibility of conducting "bottom-up" population estimation. Utilising a statistical model enables those estimates to be accompanied by estimates of uncertainty, based on their posterior marginal distributions, while explicit consideration of spatial structure can further explain spatially structured variation between surveyed areas that is not fully explained by the covariates.

Leasure et al. (2020) firstly implemented a bottom-up approach, and it used a Bayesian model with nested random effects hierarchical at different administrative levels and settlement types. The response, "microcensus" data, came from limited household surveys in randomly selected areas, while the geospatial covariates were obtained from the existing databases, including gridded population estimates from WorldPop Global High Resolution Population Denominators (WorldPop), school density from eHealth Africa, household sizes from Demographic Health Survey, settled areas and settlement types from high-resolution satellite imagery. However, the derived population estimates were not very precise. Boo et al. (2022) replaced the random sampling with a weighted one and fixed the bias brought by the weighted sampling with a weighted precision. The age and sex proportions were also modelled with the "microcensus" data aggregated at a provincial level as a Dirichlet-multinomial process. Instead of using data

from existing databases that may be available for somewhere but not for the others, they used covariates that can only be derived from the paid and proprietary techniques (i.e., the 5m resolution satellite imagery provided by Maxar Technologies and the Artificial Intelligence (AI)-based feature extraction provided by Ecopia Tech to be used to derive several characteristics of building footprints).

The "bottom-up" methods are surely good supplements to the traditional "top-down" methods in terms of providing up-to-date population estimates, but implementing the "microcensus" household surveys would cost a lot of money and several days or weeks for field works. Whether such an investment is meaningful for mapping at the required resolution still remains a question. Using ancillary data that can be easily obtained would also help reduce costs.

## 1.3 Research Purpose

This master's thesis aims to develop an approach that projects existing gridded population estimates available for free to any target combination of spatial and temporal resolutions. If the projected "top-down" population estimates based on the existing gridded population estimates available for free hold the same or even better quality than the ones achieved in the "bottom-up" approach based on the limited small-scale "microcensus", a "free lunch" can be enjoyed. That means, under such a circumstance, the user does not have to pay for the "microcensus", as paying extra does not bring about better predictive performances. This idea is the first attempt trying to assess the border between "top-down" and "bottom-up" methods, so as to find the minimum sample size that lets the predictive performance of "bottom-up" estimates be better than the projected globe population data products available for free. This idea can be applied to any gridded population estimates, including but not limited to the above-mentioned large-scale gridded population data products, so as to cover any territory of the globe. The reliability of such input population data products has yet been confirmed by a few case studies (e.g., Archila Bustos et al., 2020; Mohanty and Simonovic, 2021).

The core idea for implementing the spatial and temporal refinement is to use Bayesian hierarchical spatial and spatiotemporal models with Stochastic Partial Differential Equations (SPDE) in Integrated Nested Laplace Approximation (INLA), which have recently been accepted as useful spatial and spatiotemporal techniques in ecology and epidemiology (e.g., Adde et al., 2020; Forlani et al., 2020; Haug et al., 2020; Wilson and Wakefield, 2021). The Bayesian nature allows the user of the approach to improve the predictive performances of the "bottom-up" models based on limited survey data, by incorporating existing "top-down" gridded population estimates based on census data and available for free as the prior information. That means the proposed approach could also be regarded as a combination of "top-down" and "bottom-up" approaches.

Nighttime Lights (NTL) on the earth's surface have often been used as an indicator of the spatial distribution of population in recently proposed "top-down" approaches with dasymetric mapping (e.g., Bagan and Yamagata, 2015; Sun et al., 2017; Wang et al., 2018; Yu et al., 2018; Lu et al., 2021), as well as the ones already contributing to large-scale gridded population data products (e.g., Balk et al., 2005; Lloyd et al., 2019). This is because the strong correlation between NTL and population has been observed in most countries of the globe (Elvidge et al., 2014). Stathakis and Baltas (2018) estimated seasonally specific ambient population counts at sub-national level to reflect the substantially altered demand for social services, thanks to

a proportional relationship and a high refresh rate of NTL data.

Simplifying the use of ancillary data makes projecting existing gridded population estimates becomes possible, because the user would not have to collect a large amount of ancillary data against certain standards. Ancillary geospatial data used in the proposed models include NTL data available sub-yearly and publicly, as well as the easy-to-update or hard-to-change ones usually used in a comprehensive "top-down" population mapping approach (e.g., digital elevation). The other easy-to-change but hard-to-update ancillary data (e.g, road networks) are no longer employed, but the spatial and temporal uncertainties caused by the simplification are simultaneously better concluded by the hierarchical spatial and spatiotemporal models, relative to the non-spatial models. This is of great importance for decision-makers. For instance, a vaccination campaign would be interested in the upper bound to allocate sufficient amount of resources to target population, while a telecommunication company would be interested in the lower bound to deploy base stations in a cost-effective way.

It is the first time to use data available for free for both covariates and "top-down" population data products as the prior information for "bottom-up" estimates. The covariates used by Leasure et al. (2020) and Boo et al. (2022) cannot be employed to derive such prior information at a low cost, because deriving such information requires data for covariates covering the whole target region. The ones proposed by them can only be acquired from existing databases or the paid and proprietary techniques. It is also the first time to project data available only at a coarser resolution (bad global mapping) to a finer one (target regional mapping) and meanwhile extend the model to a temporary one (projecting yearly population with NTL data available monthly and for free). NTL were photoed globally for the same time span with same quality which would be better than the satellite imagery of building footprints provided by Maxar Technologies which possibly contains very outdated data (see Boo et al. (2022) who used image on average newer than 2017 while the worst one produced in 2009).

Obviously, the proposed refining approach is not able to improve the quality of original large-scale population data products, but to improve the spatial and temporal availability of them. For instance, if a user would like to investigate population distribution during holiday travel seasons, Oak Ridge National Laboratory's LandScan dataset would be the sole candidate focusing on the large-scale daytime population estimation, but only available yearly at 30 arc-seconds. Naturally, the user may wish to reproduce the LandScan data at the resolutions required by the intended application, but the algorithms used by the LandScan data are proprietary. The situation urges the use of approach proposed in this master's thesis to tailor the original population data products to the user's need, so other kinds of information do not have to be aggregated to coarser spatial and temporal levels and thus no information would be lost. The associations with covariates should not be interpreted causally, which may result in simplistic environmental determinism arguments (Wardrop et al., 2018).

# Chapter 2

# Data

This chapter introduces three types of large-scale gridded population data to be used in the following chapters, with two of them treated as the real population data to be sampled as the "microcensus" data and the remaining one treated as the existing population data to be projected and the precise and less precise estimates on the other two types. It also introduces the ancillary data to be used in the following chapters, the specific reasons of not using some commonly used ancillary data (i.e., the idea of simplifying the use of ancillary data), the data processing platforms and the preprocessing procedures.

## 2.1 Large-Scale Gridded Population Data

In Europe, only several countries have ground-truth demographic data available at a resolution of 1 km or finer (Batista e Silva et al., 2021), while the Netherlands is the only country whose georeferenced demographic data contain the information of school density and average household size per cell. Such information allows a comparison between the approach proposed in this master's thesis and the "bottom-up" approach implemented in Leasure et al. (2020). The statistical data per cell and zip code 2015 provided by the Centraal Bureau voor de Statistiek (CBS) of the Netherlands are available at resolutions of 100 m and 500 m and derived from the Basic Registration of Persons (BRP) and the Basic Registration of Addresses and Buildings (BAG), the Valuation of Immovable Property (WOZ) register, the Geographical Basic Register (GBR) and the Integrated Income and Asset Statistics (IIVS) (CBS, 2019).

However, for reason of reliability and confidentiality, the CBS discarded the population counts below 5 and round the remaining ones by 5, while the regional population totals provided by the CBS possibly outnumber the ones recorded in the Dutch national census (see the exploratory data analysis section in Chapter 5). Hence, the refining approach was chosen to conduct at 15 arc-seconds level (i.e., to project original population estimates available at a resolution of 1 km to 500 m) in order to minimise the uncertainty brought by this artificial ambiguity. To facilitate the test of the idea of assessing the border between "top-down" and "bottom-up" approaches, the CBS georeferenced population counts were once treated as the real population counts to be sampled in Chapter 5.

For Chapter 4 and Chapter 5, the gridded population estimates contained in Global Human Settlement Population Grid (GHS-POP) version R2019A were once accessed, which depict the population distribution and density, expressed as population count per cell. The data package disaggregated residential population estimates for the target years 1975, 1990, 2000

and 2015 collected from Gridded Population of the World (GPW) version 4.10 into 9 and 30 arc-seconds grid cells (approximately 250 m and 1 km at the Equator) on the basis of the distribution and density of built-up as mapped in Global Human Settlement Layer (GHSL) global layer (Schiavina et al., 2019). In Chapter 4 and Chapter 5, the CBS population data-alike GHS-POP data at a 30 arc-seconds resolution were used as the prior information and tested as the data to be projected, because the redistributed census-based de jure/de facto population counts are constrained within the built-up regions. The GHS-POP data at a 15 arc-seconds resolution (to be aggregated from a 9 arc-seconds resolution in advance), based on which the coarser ones were aggregated, were once treated as the real population counts to be sampled in Chapter 4. It is because the GHS-POP data at a 30 arc-seconds resolution are actually less precise estimates on the CBS population data a 15 arc-seconds resolution, and it is interesting to know how the border between "top-down" and "bottom-up" approaches will look like, if the estimation errors contained in the original gridded population data products are removed (i.e., the GHS-POP data at a 30 arc-seconds resolution are actually precise estimates on the GHS-POP data a 15 arc-seconds resolution).

For Chapter 6, the unconstrained gridded population estimates produced by WorldPop were accessed, which involve population counts per pixel with country totals adjusted to match the corresponding official United Nations population estimates (WorldPop, 2022). The WorldPop programme used census-based population inputs from GPW version 4 to produce a time-consistent series of yearly population estimates from 2000 to 2020, within 3 and 30 arc-seconds grid cells (approximately 100 m and 1 km at the Equator) with the former ones aggregated to the latter ones. Its disaggregation approach used country-specific Random Forest dasymetric classifications to create a predictive weighting layer based on a variety of ancillary covariate layers including NTL data and slope derived from Digital Elevation Models (DEM) (Stevens et al., 2015; Lloyd et al., 2019). This is because the GHSL programme did not provide such a rich time series, while the constrained WorldPop data were only available for 2020.

## 2.2 Geographic Ancillary Data

Stevens et al. (2015) criticised the large amount of ancillary data used by WorldPop programme for the difficulty in standardisation and the presence of many high correlations and nonlinear interactions. Following the idea of simplifying the use of ancillary data firstly mentioned in the research purpose section of Chapter 1, several types of geographic ancillary data were discarded: (i) roads, waterways, settlements, protected areas and facilities, as they are easy to change but hard to accurately investigate sub-yearly especially in developing countries; (ii) mean temperature and precipitation, since quality of such data is not equally high in all places due to non-uniform climate station density (Hijmans et al., 2005) and that may lead to unaccountable geographically inconsistent uncertainty; (iii) comprehensively constructed land cover, land use and urbanisation, because reproducing these data for target time and regions through the comprehensive algorithms may introduce unaccountable compound uncertainty. Finally, the frequently updated spatially and temporally consistent NTL data and the time-invariant DEM were reserved in the simplification.

NTL data from Defense Meteorological Satellite Program (DMSP)-Operational Linescan System (OLS) and Visible Infrared Imaging Radiometer Suite (VIIRS)-Day Night Band (DNB) on Suomi National Polar-orbiting Partnership satellite offer a great chance of dynamically moni-

toring human activities on a large scale. A global NTL dataset produced by Li et al. (2020) was accessed, which harmonises the inter-calibrated NTL observations from DMSP-OLS version 4 data and the simulated DMSP-like NTL observations from VIIRS-DNB data. The integrated and consistent series from 1992 to 2018 quantifies NTL strength with Digital Number (DN) values ranging from 0 to 63 and a spatial resolution of 30 arc-seconds. Slope derived from Multi-Error-Removed Improved-Terrain DEM (MERIT DEM) was another ancillary data used rather than elevation itself, because population may exist in plateau where elevation is high but still habitable (Sun et al., 2017). MERIT DEM produced by Yamazaki et al. (2017) represents terrain elevations in metre at a 3 arc-seconds resolution and covers land areas between 90N-60S, while slope is defined by non-negative numbers less than 90.

Besides, information of gemeente-, provincie- and country-level administrative borders of the Netherlands produced by CBS (2015) and that of county-level administrative borders of Kenya produced by Ndeng'e et al. (2003) were accessed to tailor population and ancillary data to the models' needs. The three hierarchies of administrative levels of the Netherlands enable flexible ways of extracting small amounts of data to satisfy the implementation of research designs at low computation costs, while the data extracted from Nairobi may witness the fast urbanisation in a developing country.

## 2.3 Data Processing Platforms

The preliminary geographic data processing was completed with QGIS version 3.20.0 (QGIS Development Team, 2022). The statistical analysis was executed with R version 4.1.2 (R Core Team, 2022) in RStudio Server version 2022.2.0.443 (RStudio Team, 2022) on two Intel® Xeon™ Platinum 8163 virtual CPUs @ 2.5 GHz and 16 GiB RAM.

## 2.4 Data Preprocessing Procedures

All geographic data used in this thesis were initially acquired in or finally adjusted to coordinate reference system WGS84. The population-related and ancillary data had to be aligned at first, as they were obtained from different sources. All the data were aligned at a resolution of roughly 30 or 15 arc-seconds with the nearest neighbour algorithm, depending on the analyses. Zero was assigned to wherever a value was not available.

# Chapter 3

# Methods

The statistical methods introduced in this chapter is the basis of implementing the idea of projecting existing gridded population estimates and the idea of using existing "top-down" gridded population estimates as prior information for the "bottom-up" models with limited "microcensus" survey data. Details of how to use them are presented in the research design sections of the next two chapters.

## 3.1 Basic Spatial Modelling

As an observation is more correlated with an observation closer in space, the spatial models can be regarded as models with correlated random effects. All the input data are well gridded and thus could be regarded as point-referenced data and also observations of a spatial process $Y(\boldsymbol{s})$, population counts in this master's thesis, defined as Equation $(3.1)$

$$Y(\boldsymbol{s}) \equiv \{y(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{D}\} \tag{3.1}$$

with $\boldsymbol{s}$ being a spatial index which varies continuously in domain $\mathcal{D}$, a fixed subset of $\mathbb{R}^d$. Observations available at $n$ spatial locations are denoted by the vector $\boldsymbol{y} = (y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_n)) = (y_1, ..., y_n)$.

$Y(\boldsymbol{s})$ is a Gaussian Field (GF), since it is assumed that, for any $n \geq 1$ and for each set of locations $\boldsymbol{s}_i$, $\boldsymbol{y}$ follows a multivariate Normal distribution with mean $\boldsymbol{\mu}$ and spatially structured covariance matrix $\boldsymbol{\Sigma}$. Generic element of $\boldsymbol{\Sigma}$ is defined by a covariance function $\mathcal{C}$, such that $\boldsymbol{\Sigma}_y = \text{Cov}(y_i, y_j) = \mathcal{C}(y_i, y_j)$. As $Y(\boldsymbol{s})$ is assumed to be second-order stationary, mean function is constant in space (i.e., $\mu(\boldsymbol{s}_i) = \mu$ for each $i$) and spatial covariance function depends only on distance vector $(\boldsymbol{s}_i - \boldsymbol{s}_j) \in \mathbb{R}^2$ (i.e., $\text{Cov}(y_i, y_j) = \mathcal{C}(\boldsymbol{s}_i - \boldsymbol{s}_j)$). $Y(\boldsymbol{s})$ is further assumed to be isotropic, and that means covariance function no longer depends on direction but only on Euclidean distance $\|\boldsymbol{s}_i - \boldsymbol{s}_j\| \in \mathbb{R}$.

It is assumed that there is a latent GF $\xi(\boldsymbol{s}_i) = \xi_i$, a realisation of the spatial process, that cannot be directly observed. Instead, observations are data with a measurement error $e_i$, as presented in Equation $(3.2)$.

$$y_i = \xi_i + e_i \tag{3.2}$$

It is assumed that $e_i$ is independent of $e_j$ for all $i \neq j$ and $e_i$ follows a Gaussian distribution with zero mean and variance $\sigma_e^2$. $\sigma_e^2$ is also known as the "nugget effect" in geo-statistics. Then covariance of marginal distribution of $y(\boldsymbol{s})$ at a finite number of locations is $\Sigma_y = \Sigma_\xi + \sigma_e^2 \boldsymbol{I}$.

## 3.2 Latent Gaussian Modelling

A likelihood for non-Gaussian $\boldsymbol{y}$ conditional on an unobserved random effect (i.e., a GF) is assumed and Equation $(3.2)$ is extended with a hierarchical model and further a latent Gaussian one with an additional assumption that each of unknowns follows Gaussian distribution in the linear predictor. Three types of transformation for $y_i$ are specified by defining $y_i$ as a function of a structured additive predictor $\eta_i$ through a transformation function $g$, such that $E(g(y_i)) = \eta_i$. The transformation functions tested are $g_1(y_i) = y_i$ (i.e., no transformation), $g_2(y_i) = \sqrt{y_i}$ and $g_3(y_i) = \ln(y_i + 1)$. For $g_3(y_i)$, back transformed predicted values below or equal to zero are regarded as predicted population counts of zero and therefore assigned with zero. Additive linear predictor $\eta_i$ is defined as Equation $(3.3)$, as described by Rue et al. (2009).

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \sum_{l=1}^{L} f_l(z_{li}) \tag{3.3}$$

Here $\beta_0$ is a scalar representing the intercept; the coefficients $\boldsymbol{\beta} = \{\beta_1, \beta_2\}$ qualify the (linear) effect of some covariates $\boldsymbol{x} = (x_1, x_2)$, NTL strength and slope in this master's thesis, on the response; and $\boldsymbol{f} = \{f_1, ..., f_L\}$ is a set of functions defined in terms of a set of covariates $\boldsymbol{z} = (z_1, ..., z_L)$. Different forms including spatial and spatiotemporal random effects can be assumed through terms $f_l$. All the latent (nonobservable) components of interest for inference are collected in a set of parameters, defined as $\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, \boldsymbol{f}\}$. Vector of $K$ hyperparameters is denoted as $\boldsymbol{\psi} = \{\psi_1, ..., \psi_K\}$, which expresses the knowledge on $\boldsymbol{\theta}$. Three-stage hierarchical structure of the latent Gaussian model could then be presented as follows.

By assuming conditional independence, distribution of $n$ observations is given by the likelihood, shown as Equation $(3.4)$

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{n} p(y_i|\theta_i, \boldsymbol{\psi}) \tag{3.4}$$

where each data point $y_i$ is connected to only one element $\theta_i$ in the latent field $\boldsymbol{\theta}$.

It is assumed $\boldsymbol{\theta}$ has a multivariate Normal prior with mean $\mathbf{0}$ and precision matrix $\boldsymbol{Q}(\boldsymbol{\psi})$, i.e. $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}^{-1}(\boldsymbol{\psi}))$ with density function given by Equation $(3.5)$.

$$p(\boldsymbol{\theta}|\boldsymbol{\psi}) = (2\pi)^{-n/2} |\boldsymbol{Q}(\boldsymbol{\psi})|^{1/2} \text{exp}\left(-\frac{1}{2}\boldsymbol{\theta}'\boldsymbol{Q}(\boldsymbol{\psi})\boldsymbol{\theta}\right) \tag{3.5}$$

The components of the latent Gaussian field $\boldsymbol{\theta}$ are supposed to be conditionally independent with the consequence that $\boldsymbol{Q}(\boldsymbol{\psi})$ is a sparse precision matrix, and then $\boldsymbol{\theta}$ is a Gaussian Markov Random Field (GMRF). The joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ is given by the product of the likelihood Equation $(3.4)$, of the GMRF density Equation $(3.5)$ and of the hyperparameter prior distribution $p(\boldsymbol{\psi})$, shown as Equation $(3.6)$.

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{y}) &\propto p(\boldsymbol{\psi}) \times p(\boldsymbol{\theta}|\boldsymbol{\psi}) \times p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\psi}) \\ &\propto p(\boldsymbol{\psi}) \times p(\boldsymbol{\theta}|\boldsymbol{\psi}) \times \prod_{i=1}^{n} p(y_i|\theta_i, \boldsymbol{\psi}) \end{aligned} \tag{3.6}$$

It is computationally easier to make Bayesian inference when a GMRF is used than when a GF is used (Rue and Held, 2005), because the computational cost of working with a sparse precision matrix in GMRF models is $O(n^{3/2})$ in $\mathbb{R}^2$, rather than $O(n^3)$ which causes "big $n$ problem".

## 3.3 SPDE-INLA Approach

The INLA approach, proposed by Rue et al. (2009), is regarded as an alternative to traditional Markov Chain Monte Carlo (MCMC) methods for approximating Bayesian inference. It aims to approximate the posterior marginals of the model effects $p(\theta_i|\boldsymbol{y}) = \iint p(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{y}) d\boldsymbol{\theta}_{-i} d\boldsymbol{\psi} = \int p(\boldsymbol{\psi}|\boldsymbol{y}) p(\theta_i|\boldsymbol{\psi}, \boldsymbol{y}) d\boldsymbol{\psi}$ and hyperparameters $p(\psi_k|\boldsymbol{y}) = \iint p(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{y}) d\boldsymbol{\theta} d\boldsymbol{\psi}_{-k} = \int p(\boldsymbol{\psi}|\boldsymbol{y}) d\boldsymbol{\psi}_{-k}$, by exploiting the computational properties of GMRF and the Laplace approximation for multidimensional integration.

The SPDE approach, proposed by Lindgren et al. (2011), consists in representing a continuous spatial process (i.e., a GF), by using a discretely indexed spatial random process (i.e., a GMRF). At first, a linear fractional SPDE is shown as Equation $(3.7)$

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau\xi(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}) \tag{3.7}$$

where $\Delta$ is Laplacian, $\alpha \in [0, 2]$ controls smoothness, $\kappa > 0$ is scale, $\tau$ controls variance, and $\mathcal{W}(\boldsymbol{s})$ is a Gaussian spatial white noise process.

Exact and stationary solution to this SPDE is the stationary GF $\xi(\boldsymbol{s})$ with Matérn covariance function given by Equation $(3.8)$

$$\text{Cov}(\xi_i, \xi_j) = \frac{\sigma^2}{\Gamma(\lambda)2^{\lambda-1}} (\kappa\|\boldsymbol{s}_i - \boldsymbol{s}_j\|)^{\lambda} K_{\lambda}(\kappa\|\boldsymbol{s}_i - \boldsymbol{s}_j\|) \tag{3.8}$$

where $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ is the Euclidean distance between two generic locations $\boldsymbol{s}_i, \boldsymbol{s}_j \in \mathbb{R}^d$ and $\sigma^2$ is marginal spatial variance. The term $K_\lambda$ represents modified Bessel function of the second kind with order $\lambda > 0$ measuring degree of smoothness of the process and usually kept fixed due to poor identifiability. The scaling parameter $\kappa > 0$ is related to the range $r$, whose empirically derived definition is $r = \sqrt{8\lambda}/\kappa$ corresponding to the distance at which the spatial correlation is close to 0.13, for each $\lambda \geqslant 1/2$.

The link between SPDE in Equation $(3.7)$ and Matérn parameters is given by Equation $(3.8)$ involving $\lambda$ and $\sigma^2$, presented as Equation $(3.9)$ and Equation $(3.10)$.

$$\lambda = \alpha - d/2 \tag{3.9}$$

$$\sigma^2 = \frac{\Gamma(\lambda)}{\Gamma(\alpha)(4\pi)^{d/2}\kappa^{-2\lambda}\tau^2} \tag{3.10}$$

For $d = 2$, the most natural choice $\alpha = 2$ was set, according to Whittle (1954), while Lindgren et al. (2011) and Lindgren and Rue (2015) discussed the possible alternatives. In this case, the range becomes $r = \sqrt{8}/\kappa$, and the marginal spatial variance is given by $\sigma^2 = 1/(4\pi\kappa^2\tau^2)$.

The solution to the SPDE, represented by stationary and isotropic Matérn GF $\xi(\boldsymbol{s})$, is approximated using Finite Element Method (FEM) through basis functions defined on a triangulation of the domain $\mathcal{D}$, shown as Equation $(3.11)$

$$\xi(\boldsymbol{s}) = \sum_{g=1}^{G} \varphi_g(\boldsymbol{s})\tilde{\xi}_g \tag{3.11}$$

where $G$ is total number of vertices of the triangulation, $\{\varphi_g\}$ is the set of (deterministic) basis functions, and $\{\tilde{\xi}_g\}$ are zero mean Gaussian distributed weights. Basis functions $\varphi_g(\boldsymbol{s})$ are featured with 1 at vertex $g$ and 0 at all other vertices, while, using Neumann boundary conditions, the precision matrix $\boldsymbol{Q}$ for the Gaussian weight vector $\tilde{\boldsymbol{\xi}} = \{\tilde{\xi}_1, ..., \tilde{\xi}_G\}$ is manipulated to be sparse (Lindgren et al., 2011). As a result, the GMRF $\tilde{\boldsymbol{\xi}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1})$ is obtained.

## 3.4 Spatial Modelling

When $\boldsymbol{s} \in \mathcal{D} \subset \mathbb{R}^2$ $(d = 2)$, the linear predictor, defined as Equation $(3.3)$, could be rewritten as Equation $(3.12)$, which is a spatial regression

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \sum_{g=1}^{G} A_{ig} \tilde{\xi}_g \tag{3.12}$$

where $A_{ig} = \varphi_g(\boldsymbol{s}_i)$ is generic element of the sparse matrix $\boldsymbol{A}$ that maps the GMRF $\tilde{\boldsymbol{\xi}}$ from the $G$ triangulation nodes to the $n$ observation locations, and it is replaced by $A_{mg}$ for prediction with $m$ denoting the $m$ prediction locations.

In such a case, the likelihood given as Equation 3.4 should be rewritten as Equation 3.13, as the random effects are defined as a linear combination of temporal or areal values, the elements of $\boldsymbol{\theta}$.

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{n} p\left(y_i | \sum_j A_{ij}\theta_j, \boldsymbol{\psi}\right) \tag{3.13}$$

where $A_{ij}$ is the generic element of the projector matrix $\boldsymbol{A}$.

More details on the SPDE-INLA approach could be found in Blangiardo and Cameletti (2015) and Krainski et al. (2019). This approach was implemented by using the `R-INLA` package.

## 3.5 Priors

Under the Bayesian framework, choosing priors is important but difficult. A well chosen prior could stabilise the inference and improve the predictive performance, while a poorly chosen one could be a disaster. Fuglstad et al. (2019) considered choosing prior distribution even for a stationary Gaussian Random Field (GRF) controlled only by the range $r$ and marginal spatial variance $\sigma^2$ as a challenge: (i) covariance functions of spatial Gaussian processes cause the ridge in the likelihood for $r$ and $\sigma^2$ (Warnes and Ripley, 1987); (ii) there will be no consistent estimator for all the parameters in the Matérn class, if data are observed in an increasing density in a fixed domain (Zhang, 2004); (iii) exceptions (e.g., the ratio $\sigma^2/r$ in an exponential variogram) may contain only a limited amount of information about the parameters within a bounded domain, and that may lead to very different predictive distributions with GRF of the parameters contributing to the same quantity (Fuglstad et al., 2019). That argues the use of a principled prior which allows the user to include expert knowledge of the range of physically meaningful parameters in an interpretable way. A first attempt to select priors for GRFs in a principled approach is to derive the reference priors which could be too complicated for GRFs embedded in Bayesian hierarchical models.

On the basis of the Penalised Complexity (PC) prior framework developed by Simpson et al. (2017), Fuglstad et al. (2019) proposed a joint prior for the range and marginal variance of a zero-mean Matérn GRF, appropriate for hierarchical models whose separate components are combined linearly in the latent part. The joint prior is a weakly informative prior that shrinks towards a base model with infinite range and zero marginal variance through hyperparameters. It was designed to be applied to any spatial design and any observation process, to be computationally inexpensive and to have a much simpler form than the reference priors for GRFs. The joint PC prior is considered to be a good fit to the proposed spatial refining approach

which tends to incorporate small-scale prior information obtained at a finer resolution to the estimation process run with large-scale data at a coarser resolution. This is because the joint PC prior remains meaningful when predictions are made at a higher spatial resolution than the observed data or for a larger observation area (Fuglstad et al., 2019).

In the proposed approach, the PC priors were specified for the hyperparameters on the likelihood $\sigma_e$, and on the latent effects $\sigma$ and $r$, presented as Equation 3.14 - Equation 3.16

$$P(r < r_0) = \gamma \tag{3.14}$$
$$P(\sigma > \sigma_0) = \gamma \tag{3.15}$$
$$P(\sigma_e > \sigma_{e0}) = \gamma \tag{3.16}$$

where $\gamma \in (0, 1)$ is the upper or lower tail probability of the prior distribution, while $r_0$, $\sigma_0$ and $\sigma_{e0}$ are the lower, upper and upper limits for $r$, $\sigma$ and $\sigma_e$ respectively. In this master's thesis, $\gamma$ was set as 0.05, and collect values of $r_0$, $\sigma_0$ and $\sigma_{e0}$ from the equal-tailed 95% credibility intervals of $r$, $\sigma$ and $\sigma_e$ estimated in a previous analysis (see research design sections in Chapter 4 and 5). Consequently, collecting $r_0$, $\sigma_0$ and $\sigma_{e0}$ is only based on the quantile-based credible intervals, as the coverage of the 95% Highest Posterior Density (HPD) credible intervals is further away from the nominal level and more sensitive to hyperparameters when the PC prior is used (Fuglstad et al., 2019).

Alongside the PC priors, the vague priors and the informative normal priors were also tested in Chapter 4 and 5. Vague priors were specified for the parameters not affected by the hyperparameters as Gaussian distributions with zero mean and precision 0 for the intercept $\beta_0$ and 0.001 for the coefficients of the covariates $\beta_1$ and $\beta_2$, as introduced by Krainski et al. (2019). By using the simplest representations $\theta_1 = \ln(\tau(\boldsymbol{s})) = \ln(\tau)$ and $\theta_2 = \ln(\kappa(\boldsymbol{s})) = \ln(\kappa)$ described by Blangiardo and Cameletti (2015), vague priors were specified for the hyperparameters on the latent effects as $\theta_1 \sim \mathcal{N}(0, 0.1^{-1})$ and $\theta_2 \sim \mathcal{N}(0, 0.1^{-1})$, for the above-mentioned initial exploration. A vague prior was meanwhile specified for the hyperparameter on the likelihood as $\ln(\sigma_e^{-2}) \sim \ln Ga(1, 5 \cdot 10^{-5})$.

The use of normal priors with posterior mean and Standard Deviations (SD) of the hyperparameters (i.e., $\theta_1$, $\theta_2$ and $\ln(\sigma_e^{-2})$) derived from a previous analysis as mean and SD inputs was inspired by the normal-alike posterior distributions of those hyperparameters after applying some suitable transformations on $y_i$ (e.g., $g_3(y_i)$). For the coefficients of covariates $\beta_1$ and $\beta_2$, as the PC prior for this type of parameters has not been developed, the informative normal priors were specified. For the intercept $\beta_0$, a vague prior was always set. In Chapter 6, for the spatiotemporal model, a vague prior for the temporal effects was also once set as $\ln((1 + \rho)/(1 - \rho)) \sim \mathcal{N}(0, 0.15^{-1})$.

Prior settings for hyperparameters and parameters for different tests on how different types of priors behave in realising the idea of using existing "top-down" estimates as prior information and for the studies on counterpart models are summarised in Table 3.1 and applied in Chapter 4 - Chapter 6.

## 3.6 Selection Criteria

Three model checking and selection criteria were used to assess the plausibility and fit of the proposed Bayesian modelling, to determine (i) whether the projected "top-down" population

Table 3.1: Prior settings for tests on the performance of different priors in facilitating the "bottom-up" models and for studies on the counterpart models.

| Settings for | $r$ | $\sigma$ | $\sigma_e$ | $\beta_0$ | $\beta_1$, $\beta_2$... | $\rho$ | others |
|---|---|---|---|---|---|---|---|
| testing vague priors | vague | vague | vague | vague | vague | n/a | n/a |
| testing PC priors | PC | PC | PC | vague | normal | n/a | n/a |
| testing normal priors | normal | normal | normal | vague | normal | n/a | n/a |
| testing LGCP model | vague | vague | n/a | vague | vague | n/a | n/a |
| Leasure et al. (2020) | n/a | n/a | n/a | n/a | vague | n/a | vague |
| space-time extension | vague | vague | vague | vague | vague | vague | n/a |

estimates based on existing gridded population estimates available for free outperform the predictions achieved with limited "microcensus" survey data in a "bottom-up" way; (ii) the minimum sample size that allows the "bottom-up" estimates to be better than the projected "top-down" estimates; (iii) which type of priors helps improve predictive performances of the "bottom-up" models.

The leave-one-out cross-validatory Probability Integral Transform (PIT) is defined as Equation 3.17 for $\boldsymbol{y}$ coming from a continuous distribution (Dawid, 1984)

$$\mathsf{PIT}_i = p(y_i^\star \leqslant y_i | \boldsymbol{y}_{-i}) \tag{3.17}$$

while, for the discrete case, an adjusted version of the PIT can be applied (Czado et al., 2009), shown as Equation 3.18

$$\mathsf{PIT}_i^{\mathsf{adj}} = \mathsf{PIT}_i + 0.5 \times p(y_i^\star = y_i | \boldsymbol{y}_{-i}) \tag{3.18}$$

given the data omitting the observation $y_i$. A new realisation is denoted as $y_i^\star$. As suggested by Gneiting et al. (2007), if the empirical distribution of the PIT (often visualised as a histogram) is uniform, the predictive distribution will be coherent with the data. As suggested by Czado et al. (2009), U-shaped, inverse-U shaped (hump) and skewed histograms indicate underdispersed, overdispersed and biased predictive distributions respectively.

In addition, summary indices, Root Mean Square Error (RMSE) shown as Equation 3.19

$$\mathsf{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^\star)^2} \tag{3.19}$$

and Pearson Correlation Coefficient (PCC) shown as Equation 3.20, were employed to globally evaluate goodness of fit of a model

$$\mathsf{PCC} = \frac{\sum_{i=1}^{n} y_i y_i^\star - n\overline{y}\,\overline{y}^\star}{\sqrt{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}\sqrt{\sum_{i=1}^{n} y_i^{\star 2} - n\overline{y}^{\star 2}}} \tag{3.20}$$

where $\overline{y}$ and $\overline{y}^\star$ represent the means of $\boldsymbol{y}$ and $\boldsymbol{y}^\star$.

# Chapter 4

# A Theoretical Test

This chapter aims to test the idea of projecting existing gridded population estimate and the idea of assessing the border between "top-down" and "bottom-up" methods, firstly mentioned in the research purpose section of Chapter 1. It also aims to test how vague priors, weakly informative PC priors and informative normal priors, firstly mentioned in the priors section of Chapter 3, behave in realising the idea of using existing "top-down" population estimates as prior information for the "bottom-up" models with limited "microcensus" survey data. As (i) the GHS-POP data at a 15 arc-seconds resolution (aggregated from the GHS-POP data at a 9 arc-seconds resolution) were treated as unknown real population counts to be sampled and the benchmark for calculating the predictive performances; (ii) the GHS-POP data at a 30 arc-seconds resolution were treated as existing "top-down" gridded population estimates to be projected and regarded as precise estimates on the real population counts, this chapter is actually a test with precise "top-down" estimates in hand.

## 4.1 Research Design

As initially described in the research purpose section in Chapter 1, the idea of assessing the border between "top-down" and "bottom-up" methods to be tested in this chapter refers to assessing (i) whether the large-scale "top-down" population estimates well filed as the data products available for free (see Table 1.1) could be projected with the NTL and DEM data also available for free to any combination of spatial and temporary resolutions that the current databases have not yet provided; (ii) whether the projected population estimates hold the same or better quality than the ones acquired in a "bottom-up" approach, firstly mentioned in the "bottom-up" population mapping section of Chapter 1 and referring to estimating up-to-date gridded population with a statistical model, population data collected with limited household surveys in randomly selected small defined areas and several geospatial covariates available nationwide for extrapolation purpose, so the user would not have to pay for the "microcensus" or the paid and proprietary techniques (i.e., satellite imagery of building footprints and proprietary AI-based feature extraction) as used by Boo et al. (2022), or relay on the non-popularised ancillary data as used by Leasure et al. (2020). Thanks to the idea of assessing the border between "top-down" and "bottom-up" methods, the user could find the minimum sample size that allows the predictive performance of the paid "bottom-up" estimates to be better than the free projected global population data products, so the user could decide whether to invest in this partial-enumeration "microcensus" and satellite imagery, estimated to cost between

$0.03 and $0.15 per person in the population (Wardrop et al., 2018) and also several days or weeks for field works.

In this chapter, improving the spatial resolution of gridded population data from 30 arc-seconds to 15 arc-seconds is raised as an example, simply because the more reliable CBS population data presented in the Chapter 5 are only provided at a resolution of 500 m (roughly 15 arc-seconds). In a real practice, the user of the proposed approach may, for instance, downscale gridded population data at a spatial resolution of 1 km to 250 m, and it largely depends on how well the original large-scale population data products estimate the real population distributions and counts. The GHS-POP data at a 30 arc-seconds resolution for the year of 2015 were treated as existing "top-down" gridded population estimates to be projected to a finer 15 arc-seconds resolution that has never been provided by the GHSL programme. The GHS-POP data at a 9 arc-seconds resolution for 2015 were aggregated to 15 arc-seconds, treated as unknown real population counts, and used as the benchmark for calculating predictive performances of the models. As discussed in the large-scale gridded population data section of Chapter 2, the existing "top-down" estimates shown in Table 1.1 can only be regarded as estimates on the real population distributions and counts, and inevitably they contain measurement errors and statistical errors as mentioned in the "top-down" population mapping section of Chapter 1. It is very interesting to separate the uncertainties brought by the refining approach from the uncertainties brought by the original data products, and investigate how well the proposed refining approach would perform. That is why this test on the GHS-POP data at a 15 arc-seconds resolution is proposed, because the GHS-POP data product was produced at a 250 m resolution, and then aggregated at a 1 km resolution (Schiavina et al., 2019).

The analysis was conducted with all the data clipped to fit the territory of provincie of Utrecht of the Netherlands, so as to reduce computation costs. The estimation was made with data at a spatial resolution of 30 arc-seconds, while population counts were predicted on a refined 15 arc-seconds grid of the same dimensions where the NTL data for 2015 were disaggregated through the bilinear interpolation and aligned together with the other data. Basically, that is how the proposed approach functions to improve the spatial resolution of gridded population data from 30 arc-seconds to 15 arc-seconds. The reason of not meanwhile directly disaggregating population data and estimating at a 15 arc-seconds resolution is, under this setting, the extra uncertainty brought by population data disaggregation cannot be measured by the model. It has to be noted that the posterior mean of the population counts should be divided by a bridging factor (i.e., 4), since the refined cells are 4 times smaller than the original ones in areal size and, correspondingly, population counts in 15 arc-seconds cells are supposed to be 4 times smaller than the ones in 30 arc-seconds cells on average. In this case (i.e., $d = 2$), spatial models were established at the estimation locations with linear predictors derived as Equation 3.12 containing a spatial random effect $\xi_i$ which is a priori a Matérn GF, while population counts were predicted at the refined 15 arc-seconds locations, where out-of-sample predictive performance of each model was evaluated through PCC and RMSE against the 15 arc-seconds GHS-POP data to inform qualities of the projected population estimates.

As described also in the research purpose section of Chapter 1, thanks to the Bayesian nature and the easy-to-update or hard-to-change spatially and temporally consistent ancillary geospatial data, the predictive performances of the "bottom-up" models with limited "microcensus" survey data could potentially be improved with existing "top-down" gridded population estimates available for free shown in Table 1.1 as free prior information. In this chapter, this kind of combination of "top-down" and "bottom-up" approaches, which is also a combination

of census and survey data, was also assessed as a part of the idea of assessing the border between "top-down" and "bottom-up" methods. How vague priors, weakly informative PC priors and informative normal priors, introduced in the priors section of Chapter 3, behave in realising this idea of combination was also tested. By idealising the gridded population survey sampling and weighting problems in the fieldwork as suggested by Gelman (2007) and Thomson et al. (2020), it was assumed the unknown real population data could be collected in a reliable and cost-effective "microcensus" through limited household surveys in randomly selected areas. Both unweighted random sampling and random sampling weighted by the 30 arc-seconds GHS-POP data were performed in order to draw 1%, 2%, 5%, 10%, 20% and 50% of the full real population data with a fixed seed. Due to the idealisation of the sampling processes, the drawn data could be understood as either population densities or population counts, since the gridded population data were used. In a real practice, the user may consider to weight acquired survey data with the corresponding areal sizes of survey clusters of roughly the same areal sizes in order to get gridded data. To match the prior information acquired through estimating the posterior marginal distributions of parameters with the 30 arc-seconds GHS-POP data, the sampled real population data at a 15 arc-seconds resolution should be multiplied by a bridging factor (i.e., 4) before estimating with the meanwhile sampled NTL and slope data and predicting on the refined 15 arc-seconds grid with full NTL and slope data.

The test with precise "top-down" population estimates in hand, presented in this chapter, is summarised as Figure 4.1.

## 4.2 Exploratory Data Analysis

The GHS-POP data at resolutions of 30 arc-seconds (i.e., the population data to be projected) and 15 arc-seconds (i.e., the real population data in this chapter) are visualised as Figure A.1 and available in Appendix A, with the transformation function $g_3(y_i)$ applied in order to show the patterns more clearly. Summary statistics of the GHS-POP data are presented as Table 4.1. It can be found that, on average, the real population data have more small counts and fewer large counts than the ones to be projected, but the population total of the former one is larger than the later one within the territory of provincie of Utrecht.

Table 4.1: Summary statistics of the GHS-POP data.

| GHS-POP | Min | 25% | 50% | Mean | 75% | Max | Total |
|---|---|---|---|---|---|---|---|
| 30 arc-seconds | 0.00 | 24.15 | 119.21 | 466.44 | 505.29 | 3639.93 | 1268709 |
| 15 arc-seconds | 0.000 | 2.024 | 35.051 | 168.445 | 178.576 | 1278.376 | 1803032 |

Figure 4.2a shows a plot of the classic and robust versions of empirical semivariogram values for 0.003 degree wide bins of semivariograms of the GHS-POP data at resolutions of 30 arc-seconds (i.e., the population data to be projected) and 15 arc-seconds (i.e., the real population data in this chapter), and the official CBS data at a 15 arc-seconds resolution (i.e., the real population data in Chapter 5, denoted as "CBS-POP" and to be focused in Chapter 5) on the transformed scale $g_3(y_i)$. The empirical semivariograms are constructed as a measure of central tendency of squared differences in the transformed population data between pairs of

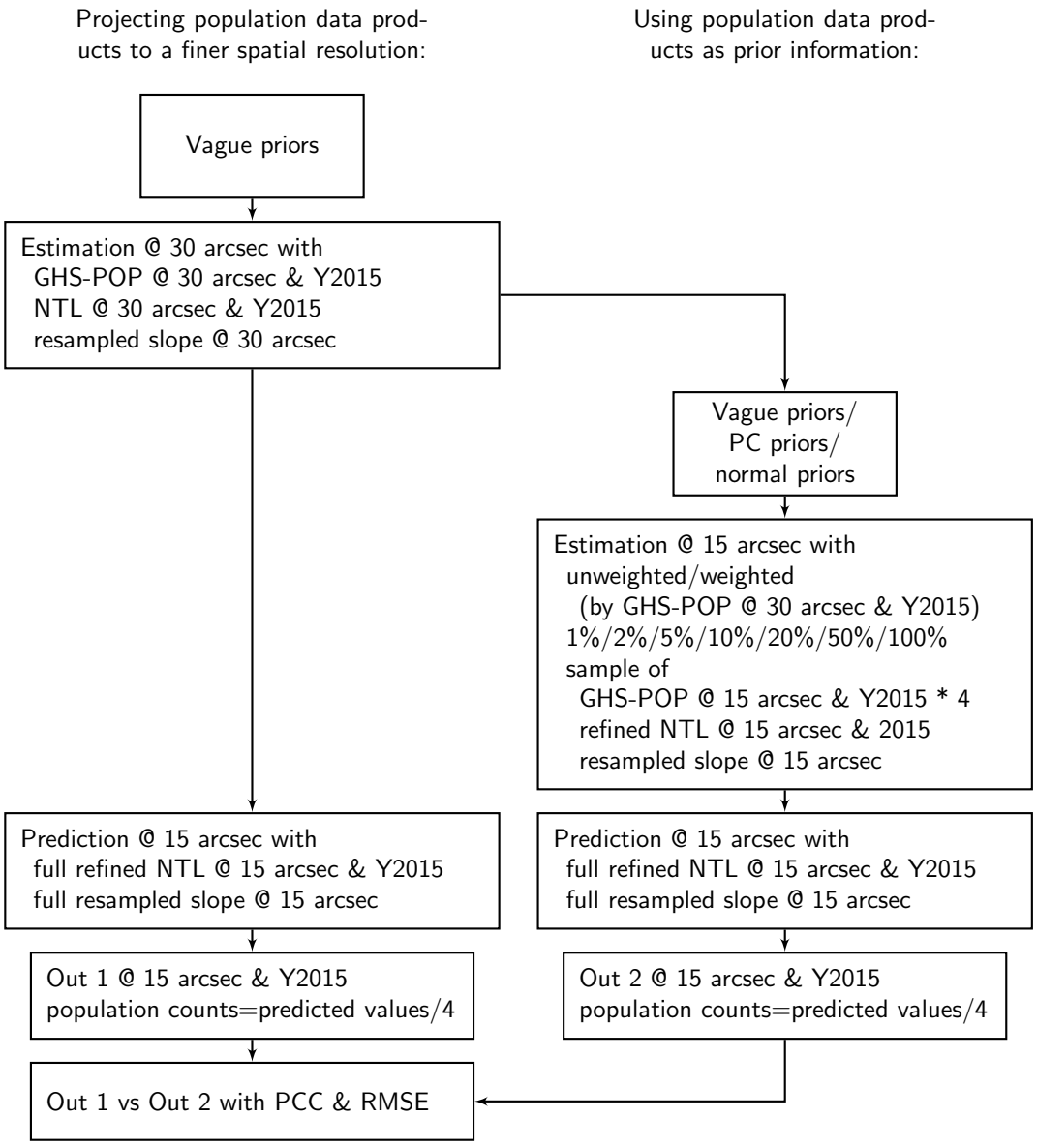



Figure 4.1: A test with precise "top-down" population estimates in hand.

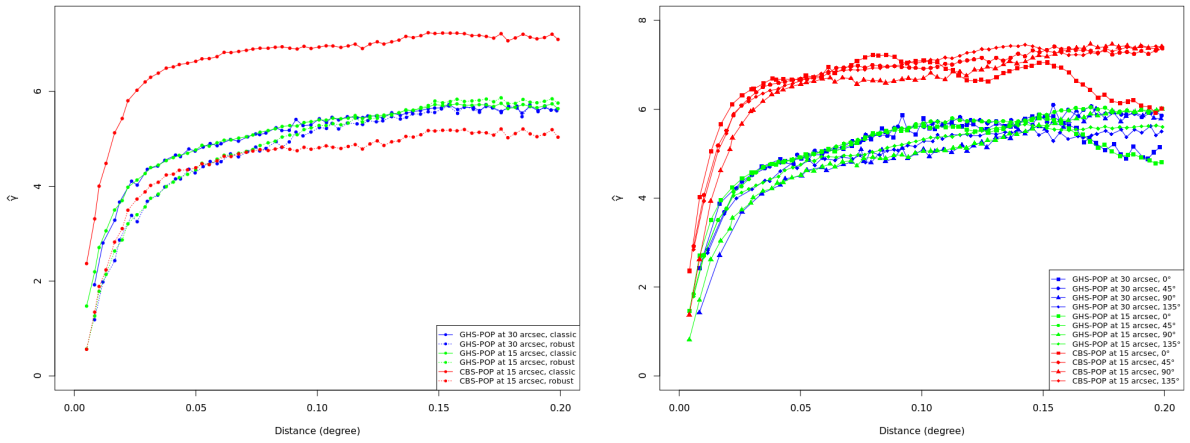
points whose inter-point distance falls into the bin (Bivand, 2010), expressed as Equation 4.1

$$\hat{\gamma}(h \pm \delta) := \frac{1}{2|N(h \pm \delta)|} \sum_{(i,j) \in N(h \pm \delta)} |g_3(y_i) - g_3(y_j)|^2 \tag{4.1}$$

where $N(h \pm \delta) \equiv \{(\boldsymbol{s}_i, \boldsymbol{s}_j) : |\boldsymbol{s}_i, \boldsymbol{s}_j| = h \pm \delta; i, j = 1, ..., N\}$ with $h$ separating each pair of points and $\delta$ representing bin width tolerance range. The robust semivariograms, developed by Cressie (1993) and expressed as Equation 4.2, reduce the influence of unusually large differences in value between near neighbours, and the large impact of differences in value between near neighbours is found (to be discussed in Chapter 5).

$$\hat{\gamma}(h \pm \delta) := \frac{\left\{ \frac{1}{|N(h \pm \delta)|} \sum_{(i,j) \in N(h \pm \delta)} |g_3(y_i) - g_3(y_j)|^{1/2} \right\}^4}{2\left(0.457 + \frac{0.494}{|N(h \pm \delta)|}\right)} \tag{4.2}$$

Generally, the assumption of second-order stationarity, mentioned in the basic spatial modelling section of Chapter 3, is found reasonable, because the semivariograms of those population data are roughly bounded (Atkinson and Lloyd, 2009).




(a) (b)

Figure 4.2: Classic and robust versions of (a) the empirical variograms and (b) the directional empirical variograms of the GHS-POP data at resolutions of 30 arc-seconds in blue (i.e., the population data to be projected) and 15 arc-seconds in green (i.e., the real population data defined in this chapter), and the official CBS data at a 15 arc-seconds resolution in red (i.e., the real population data to be used in Chapter 5, denoted as "CBS-POP") on the transformed scale $g_3(y_i)$.

Figure 4.2b shows a plot of four empirical variograms for four axes at 0°, 45°, 90°and 135°and for transformed population data from different sources, it is found that the assumption of isotropy, as mentioned in the basic spatial modelling section of Chapter 3, is also rational, since the semivariograms in four different directions look alike (Bivand et al., 2008).

## 4.3 Counterpart: LGCP Model

Considering that the population density seems to be a kind of intensity, a very natural idea is to model the average population counts per unit of space (i.e., population density) associated with an underlying spatial process (i.e., a point pattern) with an intensity function $\lambda(\boldsymbol{s})$. A Log-Gaussian Cox Process (LGCP) refers to modelling the log intensity of the Cox process with a Gaussian linear predictor under the setting of a log-Cox point process model (Møller et al., 1998), and can be expressed as Equation 4.3 and Equation 4.4 for gridded population counts.

$$y_i \sim \text{Pois}(\lambda(\boldsymbol{s}_i)) \tag{4.3}$$

$$\ln(\lambda(\boldsymbol{s}_i)) = \eta_i \tag{4.4}$$

However, given, as suggested by Thomson et al. (2020), (i) the "top-down" population data products that restrict estimates to settled area likely underestimate rural population and overestimate the urban one (e.g., GHS-POP data); (ii) the data products that estimate population in all landmasses likely overestimate rural and underestimate urban one (e.g., WorldPop data), (i) the Poisson process assumed may lead to a serious underdispersion or overdispersion problem; (ii) how severe the exact dispersion problem is largely depends on the modelling process of input population data, the target region of interest and its areal size. The importance of using a Gaussian process with a dispersion parameter $\sigma_e^2$ instead in modelling population counts is stressed in the results section of this chapter.

## 4.4 Results

Figure 4.2a suggests the ranges of the semivariograms were around 0.025 degree for all those population data, which means 0.025 degree is the distance in which the difference of the variogram from the sill becomes negligible. As introduced in the SPDE-INLA approach section of Chapter 3, the SPDE approach is based on a triangulation of the spatial domain. Blangiardo and Cameletti (2015) stated that number of vertices used in the triangulation was determined by a trade-off between accuracy of the GMRF representation and computation costs, so it is natural to consider a sub-0.025-degree largest allowed triangle edge length. The final chosen largest allowed triangle edge lengths are 0.012 degree for the inner area which covers the whole territory of interest, and 0.03 degree for the outer area which is an extension of the original domain in case of the boundary effects related to the SPDE approach as suggested by Lindgren and Rue (2015). A cut-off on edge length of 0.01 degree is used to avoid building too many small triangles around some certain locations. The triangular mesh used for all the calculations, mentioned in Chapter 4 and Chapter 5, can then be visualised as Figure 4.3.

As introduced in the research design section of this chapter, the tests on whether existing "top-down" gridded population estimates could be used as prior information so as to improve the predictive performances of the "bottom-up" models require limited "microcensus" survey data to be collected randomly in both unweighted and weighted ways with the collected sample sizes equal to 1%, 2%, 5%, 10%, 20% and 50% of the real population data. The final collected samples are visualised as Figure 4.4, with larger dots firstly collected in a sample with a smaller sample size due to the use of a fixed seed which allows the previously collected data to be collected again.

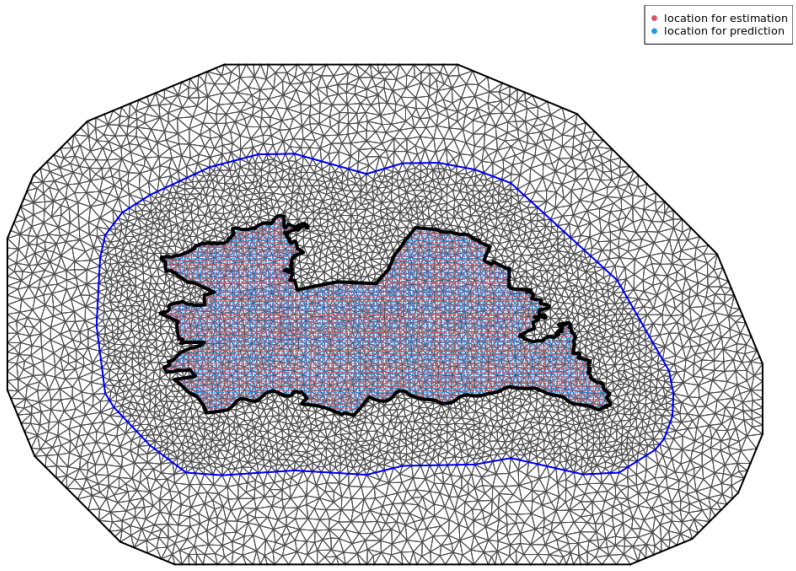



Figure 4.3: Triangular mesh used for spatial modelling with estimation locations (i.e., the locations where the GHS-POP data at a 30 arc-seconds resolution laid) indicated by red dots and prediction locations (i.e., the locations where the refined 15 arc-seconds grid was located by densifying where the GHS-POP data at a 30 arc-seconds resolution laid 4 times) indicated by blue dots.

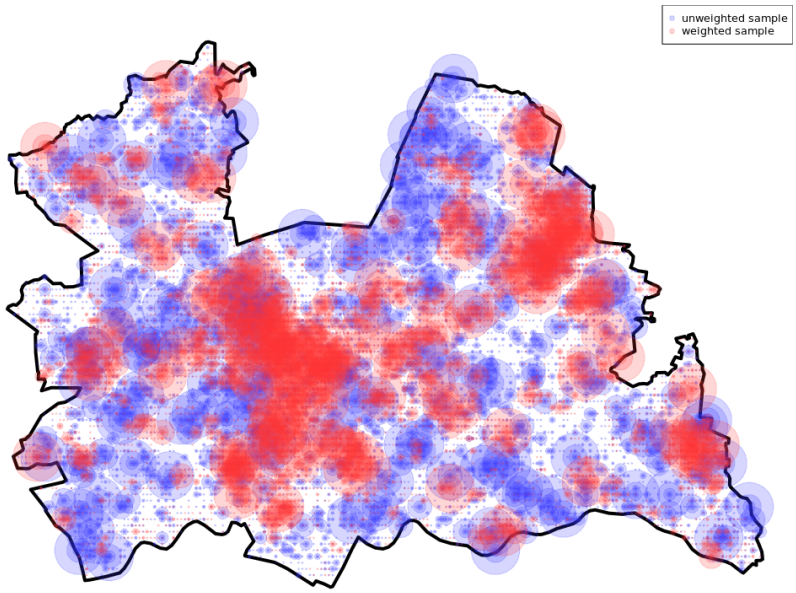



Figure 4.4: Unweighted samples, indicated by blue dots, and weighted samples (i.e., the samples collected with the 30 arc-seconds GHS-POP data as sampling weights), indicated by red dots, collected as "microcensus" survey data for testing the idea of combining the "top-down" and "bottom-up" approaches, with larger dots firstly collected in a sample with a smaller sample size (due to the use of a fixed seed).

The first attention is paid to selecting transformation functions proposed in the latent Gaussian modelling section of Chapter 3, a fundamental idea for realising the idea of assessing the border between "top-down" and "bottom-up" methods and the idea of combining the "top-down" and "bottom-up" approaches. Three estimations were made with the GHS-POP data after three types of transformations at a 30 arc-seconds resolution and associated NTL and slope data, while vague priors were set for each hyperparameters (i.e., $\theta_1$, $\theta_2$ and $\ln(\sigma_e^{-2})$) on internal scale used by `R-INLA` package and parameters (i.e., $\beta_0$, $\beta_1$ and $\beta_2$) for this initial exploration. The posterior distributions of the hyperparameters (i.e., $\sigma_e^2$, $\sigma^2$ and $r$) and parameters derived with the estimations are shown as Figure 4.5. The very large values of unstructured variance

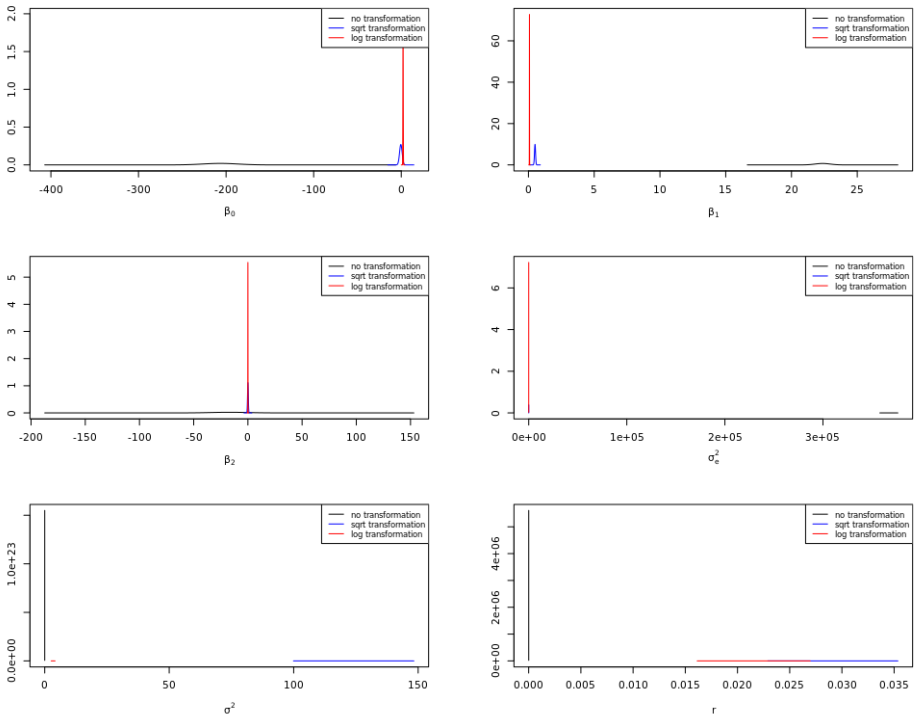



Figure 4.5: Posterior distributions of parameters and hyperparameters for estimations made with the GHS-POP data at a 30 arc-seconds resolution after three different types of transformations, $g_1(y_i)$ in black, $g_2(y_i)$ in blue and $g_3(y_i)$ in red.

$\sigma_e^2$ and structured variance $\sigma^2$ caused by transformations $g_1(y_i)$ and $g_2(y_i)$ respectively indicate the transformation $g_3(y_i)$ would be a safe option for describing spatial characteristics without overfitting.

As summarised in Table 3.1, normal priors were specified for the coefficients of the covariates $\beta_1$ and $\beta_2$ when testing whether PC priors and normal priors are good choices for realising the idea of using existing "top-down" estimates as prior information for the "bottom-up" models. When whether informative priors should be set for the coefficients of the covariates has to be confirmed in advance, whether the use of parameters $\beta_1$ and $\beta_2$ had large impacts on the posterior distributions of all the hyperparameters and parameters is of special interest. Two estimations were made with the GHS-POP data transformed with $g_3(y_i)$ at a 30 arc-seconds resolution and associated NTL and slope data included for one time and not included for another time, while vague priors were set for each hyperparameters (i.e., $\theta_1$, $\theta_2$ and $\ln(\sigma_e^{-2})$) on internal scale used by `R-INLA` package and parameters (i.e., $\beta_0$, $\beta_1$ and $\beta_2$) for this exploration.

The posterior distributions of the hyperparameters (i.e., $\sigma_e^2$, $\sigma^2$ and $r$) and parameters derived with the estimations are shown as Figure 4.6. It is found that the use of parameters $\beta_1$ and

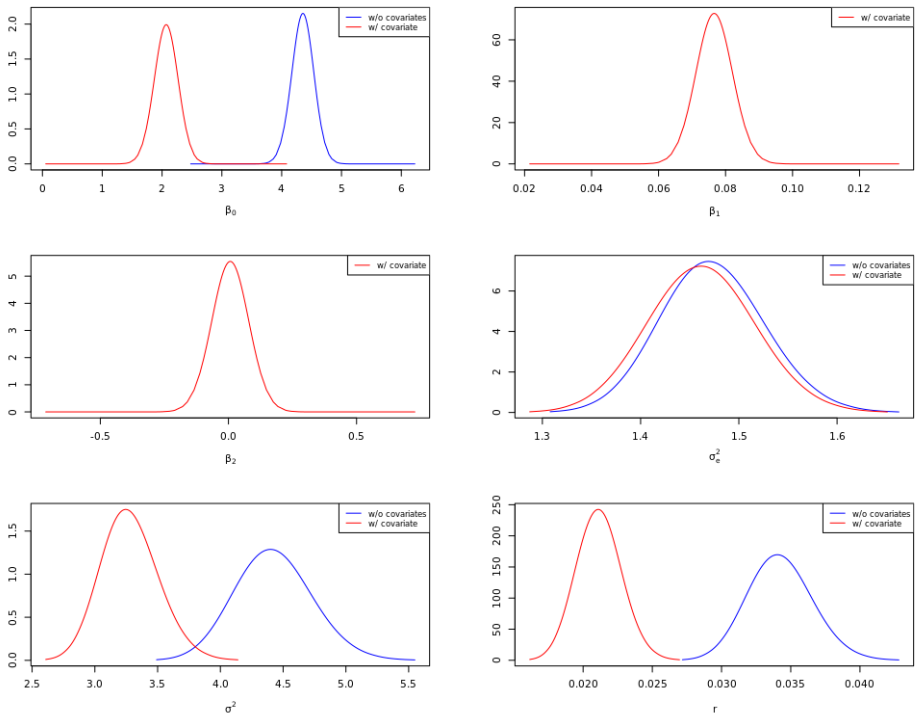



Figure 4.6: Posterior distributions of parameters and hyperparameters for estimations made with the GHS-POP data transformed with $g_3(y_i)$ at a 30 arc-seconds resolution and associated NTL and slope data included in red and not included in blue.

$\beta_2$ indeed affects the posterior distributions of all the hyperparameters and parameters, so specifying informative priors for the coefficients of covariates is considered as an appropriate choice for testing the idea of using existing "top-down" estimates as prior information for the "bottom-up" models.

Before testing whether using existing "top-down" estimates as prior information for the "bottom-up" models is a good idea, two questions have to be answered: (i) how different types of priors behave in realising the idea of using existing "top-down" estimates as prior information, and (ii) whether treating the newly sampled "microcensus" data as the prior information and the existing "top-down" estimates as the likelihood is an better idea, since the limited amount of household surveys do not necessarily convey more information than the census-based "top-down" estimates so as to be considered as a less subjective source of information. Two extreme situations were considered: (i) for the "best data with worst prior" case, three estimations were made with all the GHS-POP data, multiplied with a bridging factor 4 and transformed with $g_3(y_i)$, at a 15 arc-seconds resolution and associated NTL and slope data, while vague priors, PC priors and normal priors were set for each hyperparameters (i.e., $\theta_1$, $\theta_2$ and $\ln(\sigma_e^{-2})$) on internal scale used by `R-INLA` package after derived with the GHS-POP data transformed with $g_3(y_i)$ at a 30 arc-seconds resolution and associated NTL and slope data; (ii) for the "worst data with best prior" case, three estimations were made with the GHS-POP data transformed with $g_3(y_i)$ at a 30 arc-seconds resolution and associated NTL and slope data, while vague priors, PC priors and normal priors were set for each hyperparameters (i.e., $\theta_1$, $\theta_2$ and $\ln(\sigma_e^{-2})$)

on internal scale used by `R-INLA` package after derived with all the GHS-POP data, multiplied with a bridging factor 4 and transformed with $g_3(y_i)$, at a 15 arc-seconds resolution and associated NTL and slope data. Corresponding priors for parameters (i.e., $\beta_0$, $\beta_1$ and $\beta_2$) were set as described in Table 3.1. The posterior distributions of the hyperparameters (i.e., $\sigma_e^2$, $\sigma^2$ and $r$) and parameters derived with the estimations are shown as Figure 4.7. It is found that, the

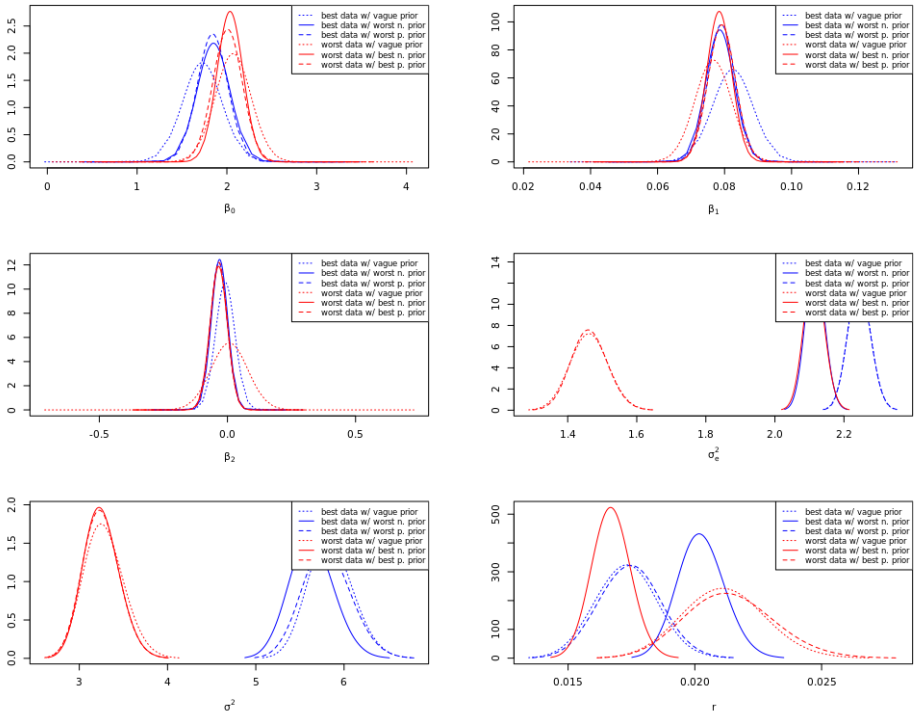
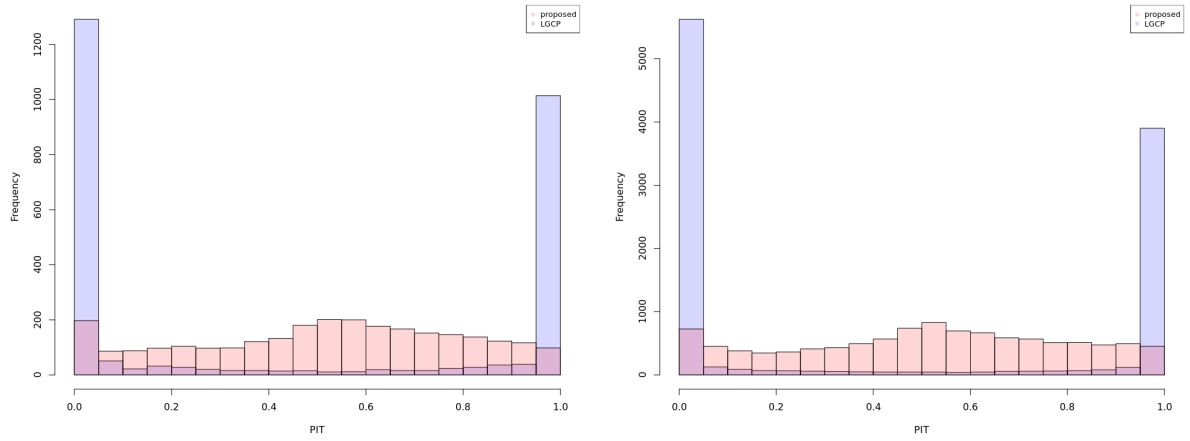



Figure 4.7: Posterior distributions of parameters and hyperparameters for estimations made with the "best data with worst prior" case in blue and the "worst data with best prior" case in red.

"peaks" of posterior distributions do reflect different levels of informativeness that different types of priors could hold, and generally the information that PC priors and normal priors conveyed does result in reduced variances of posterior distributions, with the ones associated with PC priors being more conservative. At the current stage, there has been no concrete conclusion can be made on whether the "best data with worst prior" case or the "worst prior with best data" case performs better by analysing the "peaks", since amount of information contained in the GHS-POP data at a 15 arc-seconds resolution is just four times more than that in the GHS-POP data at a 30 arc-seconds resolution. The "best data with worst prior" case and the "worst prior with best data" case brought different posterior distributions of hyperparameters and parameters, but the difference between the prior and the likelihood could be necessarily owing to variation that is not captured by the prior or likelihood alone (van de Schoot et al., 2021). This issue can be identified through a series of sensitivity analyses of the likelihoods, by examining different forms of combining census-based and survey-based information.

In the counterpart section of this chapter, the LGCP model is mentioned as a good candidate for the proposed research questions. Two estimations were made on the basis of Equation 3.2 and transformation function $g_3(y_i)$ with the GHS-POP data at both 30 arc-seconds and 15 arc-seconds resolutions and associated NTL and slope data, while vague priors were set for

each hyperparameters (i.e., $\theta_1$, $\theta_2$ and $\ln(\sigma_e^{-2})$) on internal scale used by `R-INLA` package and parameters (i.e., $\beta_0$, $\beta_1$ and $\beta_2$). Two estimations were made on the basis of Equation 4.3 and Equation 4.4 with the GHS-POP data at both 30 arc-seconds and 15 arc-seconds resolutions and associated NTL and slope data, while vague priors were set for each hyperparameters (i.e., $\theta_1$ and $\theta_2$) on internal scale used by `R-INLA` package and parameters (i.e., $\beta_0$, $\beta_1$ and $\beta_2$) for this initial exploration. It has to be noted that the GHS-POP data are not really count data but non-negative continuous data, which have to be rounded to the nearest non-negative integers in order to assess the LGCP models. The histograms of the PIT measure for these four estimations are presented as Figure 4.8a for estimations with 30 arc-seconds data and as Figure 4.8b for estimations with 15 arc-seconds data, revealing that the LGCP models suffer an underdispersion problem and thus cannot be used for the intended research purpose.




(a) (b)

Figure 4.8: Histograms of the cross-validated PIT measure for the proposed model in red and the LGCP model in blue, derived with the GHS-POP data at (a) a 30 arc-seconds resolution (i.e., the population data to be projected) and (b) a 15 arc-seconds resolution (i.e., the real population data defined in this chapter).

By now, an intermediate summary could be made: (i) the transformation function $g_3(y_i)$ is chosen; (ii) informative priors should be set on the coefficients of covariates, $\beta_1$ and $\beta_2$; (iii) the use of PC priors and less conservative normal priors indeed helps reduce the variances of posterior distributions; (iv) the newly sampled "microcensus" data and the existing "top-down" estimates are treated as the prior information and the likelihood respectively in this chapter, but more investigations should be conducted in the future; (v) the proposed model does not suffer the underdispersion problem whereas the LGCP model does. The next step is to implement the idea of assessing the border between "top-down" and "bottom-up" methods and the idea of combining the "top-down" and "bottom-up" approaches, and test them with the real population data.

Following the procedures as introduced in the research design section of this chapter and summarised as Figure 4.1, thirty-six estimations were made with each combination of six sample sizes, two sampling methods and three types of priors enumerated, to assess the idea of using "top-down" population data products as prior information. The estimations were made with the sampled GHS-POP data, once shown as Figure 4.4, multiplied with a bridging factor 4 and

transformed with $g_3(y_i)$, at a 15 arc-seconds resolution and associated NTL and slope data, while vague priors, PC priors and normal priors were set for each hyperparameters (i.e., $\theta_1$, $\theta_2$ and $\ln(\sigma_e^{-2})$) on internal scale used by `R-INLA` package after derived with the GHS-POP data transformed with $g_3(y_i)$ at a 30 arc-seconds resolution and associated NTL and slope data. Corresponding priors for parameters (i.e., $\beta_0$, $\beta_1$ and $\beta_2$) were set again as described in Table 3.1. The posterior distributions of the hyperparameters (i.e., $\sigma_e^2$, $\sigma^2$ and $r$) and parameters derived with the estimations on the basis of (i) unweighted sampling and normal priors, (ii) unweighted sampling and PC priors, (iii) unweighted sampling and vague priors, (iv) weighted sampling and normal priors, (v) weighted sampling and PC priors, and (vi) weighted sampling and vague priors are shown as Figure 4.9, Figure 4.10, Figure 4.11, Figure 4.12, Figure 4.13

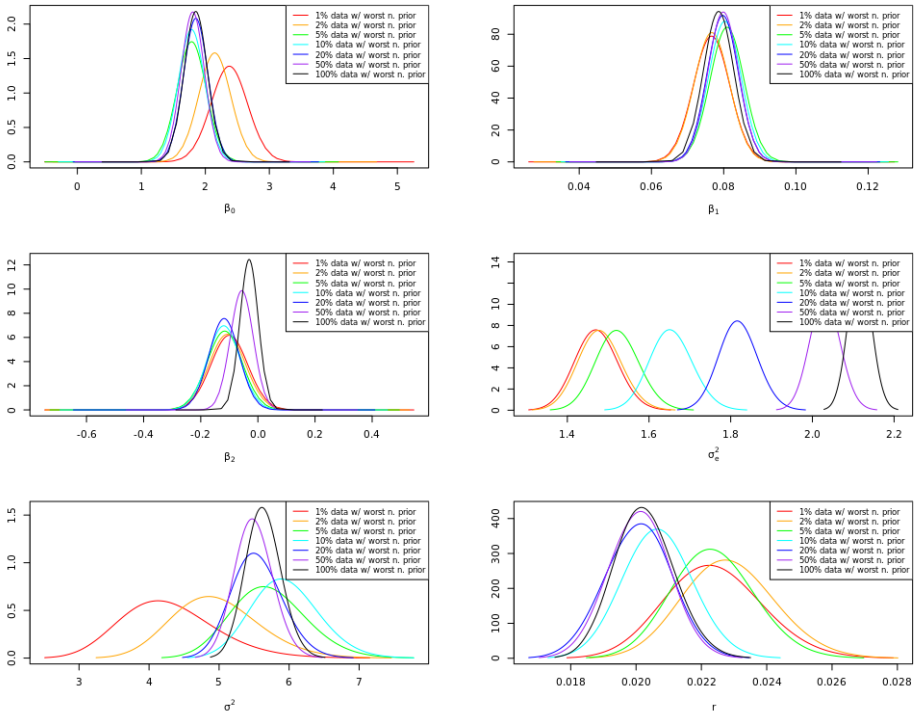



Figure 4.9: Posterior distributions of parameters and hyperparameters for estimations made with unweighted samples of 1%, 2%, 5%, 10%, 20%, 50% and 100% real population data and normal priors in red, orange, green, cyan, blue, purple and black.

and Figure 4.14. It is found that, generally, normal priors and PC priors help reduce the variances of posterior distribution of hyperparameters and parameters and could be regarded as good replacements for the vague priors, especially when the "bottom-up" estimates are based on the "microcensus" data with relatively small sample sizes. With small sample sizes, normal priors outperform PC priors, probably because the source of prior information itself (i.e., the GHS-POP data at a 30 arc-seconds resolution) could be considered as an unweighted sample of 25% the real population data (i.e., the GHS-POP data at a 15 arc-seconds resolution). Besides, PC priors function just slightly better than vague priors with small sample sizes, likely due to their mostly harmless weakly informative nature.

Table 4.2 summarises the predicted population counts on the original scale (i.e., through back transforming predicted values before assigning zero to the back transformed values below zero as mention in latent Gaussian modelling section of Chapter 3, and then dividing these values

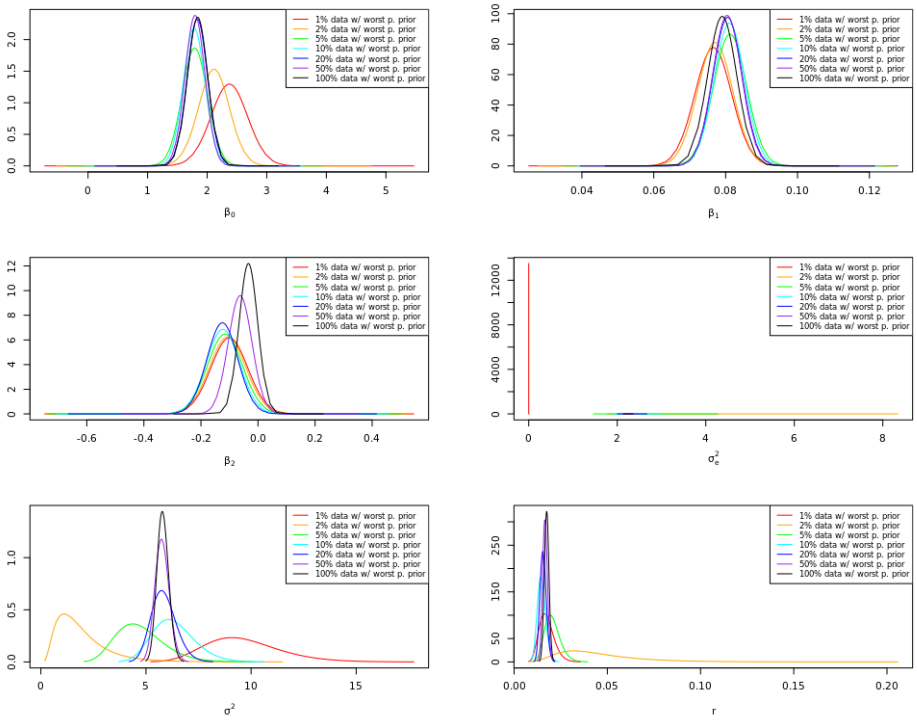



Figure 4.10: Posterior distributions of parameters and hyperparameters for estimations made with unweighted samples of 1%, 2%, 5%, 10%, 20%, 50% and 100% "fake" real population data and PC priors in red, orange, green, cyan, blue, purple and black.

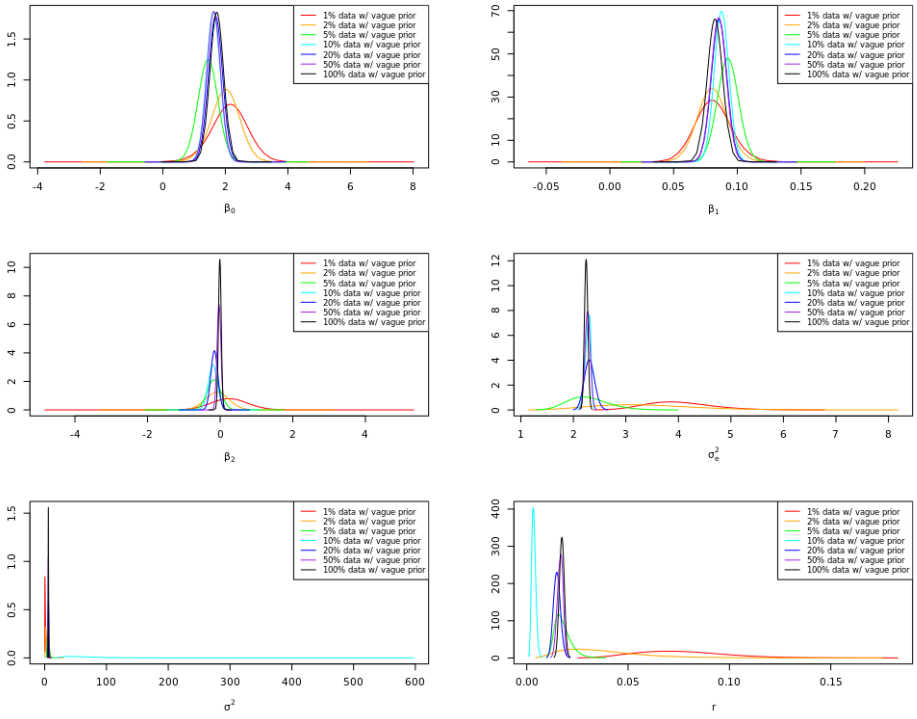



Figure 4.11: Posterior distributions of parameters and hyperparameters for estimations made with unweighted samples of 1%, 2%, 5%, 10%, 20%, 50% and 100% real population data and vague priors in red, orange, green, cyan, blue, purple and black.

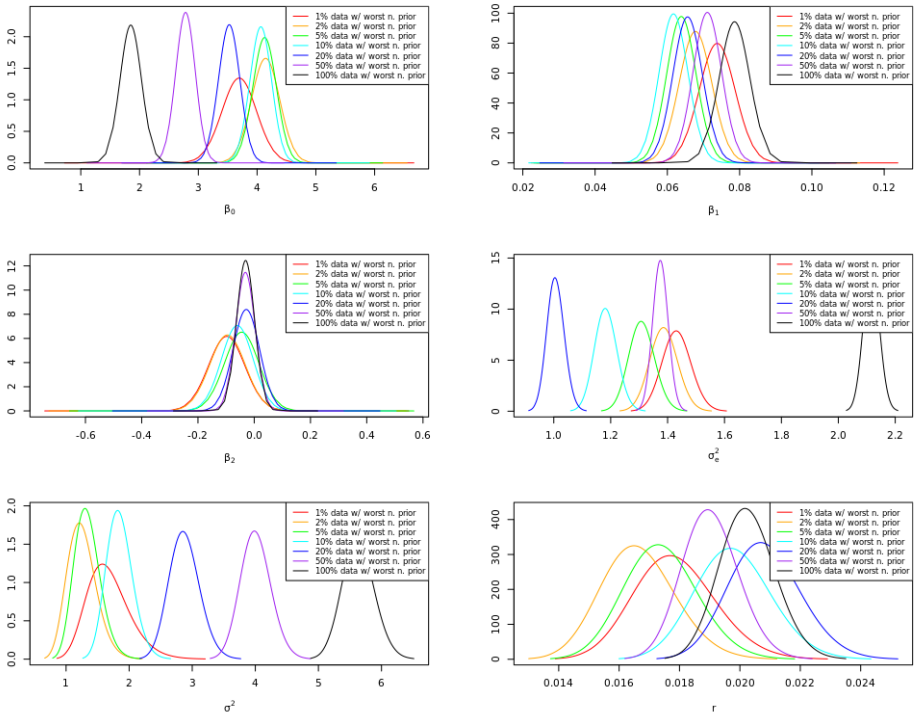



Figure 4.12: Posterior distributions of parameters and hyperparameters for estimations made with weighted samples of 1%, 2%, 5%, 10%, 20%, 50% and 100% real population data and normal priors in red, orange, green, cyan, blue, purple and black.

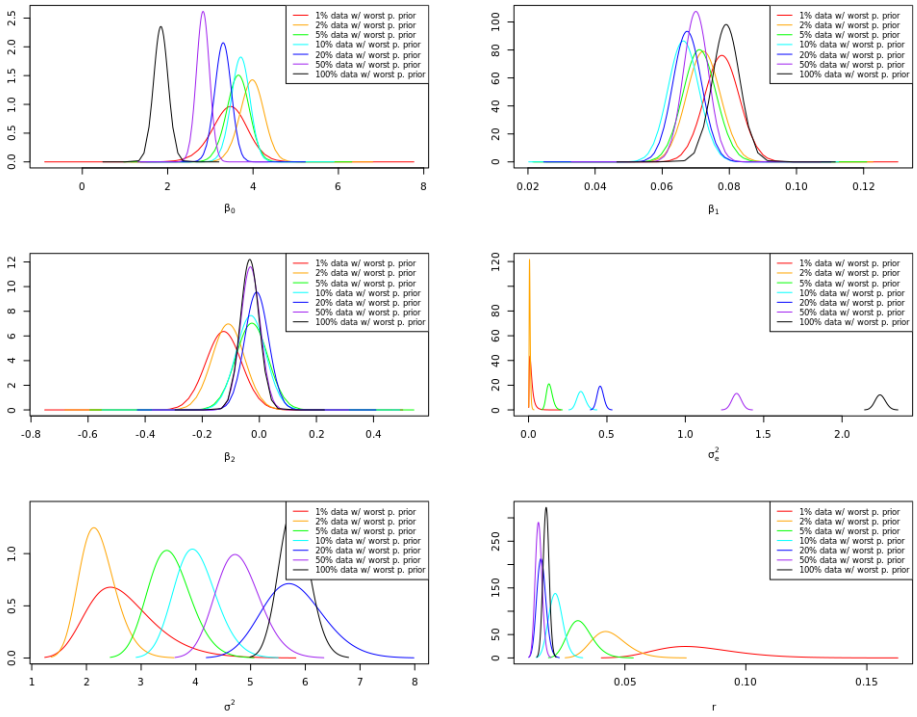



Figure 4.13: Posterior distributions of parameters and hyperparameters for estimations made with weighted samples of 1%, 2%, 5%, 10%, 20%, 50% and 100% real population data and PC priors in red, orange, green, cyan, blue, purple and black.

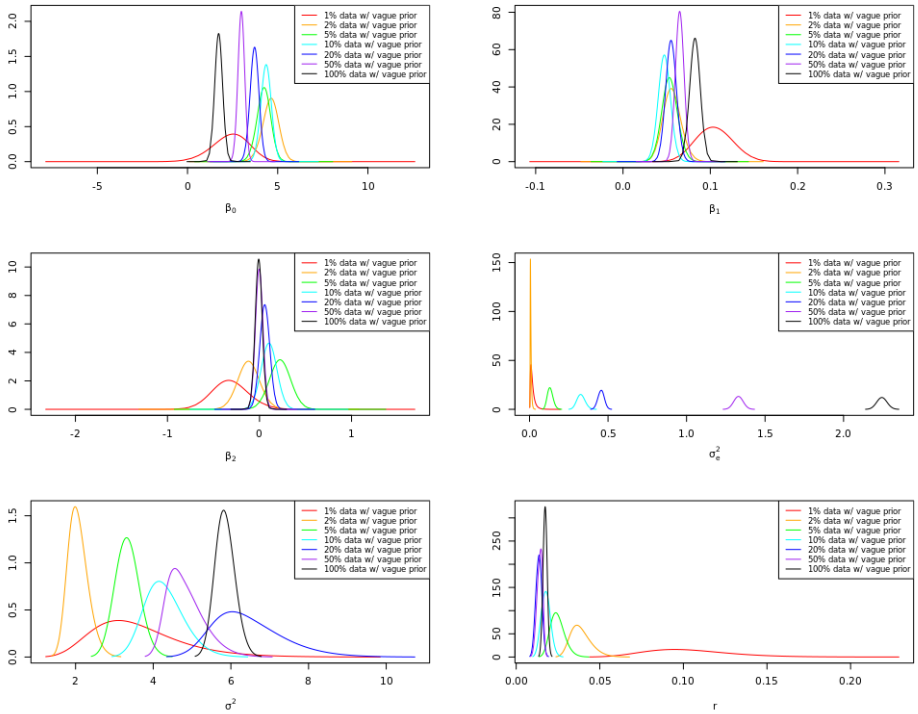



Figure 4.14: Posterior distributions of parameters and hyperparameters for estimations made with weighted samples of 1%, 2%, 5%, 10%, 20%, 50% and 100% real population data and vague priors in red, orange, green, cyan, blue, purple and black.

by 4 as shown in Figure 4.1) per 15 arc-seconds gridded cell and their totals derived on the basis of the above-mentioned thirty-six estimations, comparable to the data in Table 4.1. Figure A.2 and Figure A.3, Figure A.4 and Figure A.5, Figure A.6 and Figure A.7, Figure A.8 and Figure A.9, Figure A.10 and Figure A.5, and Figure A.12 and Figure A.13 available in Appendix A show the mapped posterior means and SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of the above-mentioned thirty-six estimations made with (i) unweighted sampling and normal priors, (ii) unweighted sampling and PC priors, (iii) unweighted sampling and vague priors, (iv) weighted sampling and normal priors, (v) weighted sampling and PC priors, and (vi) weighted sampling and vague priors, and different sample sizes. The posterior means on the transformed scale refer to back transforming predicted values before assigning zero to the back transformed values below zero, dividing these values by 4, and then transforming the quotients with $g_3(y_i)$, so as to make the results comparable to the map of the real population data shown as Figure A.1b. It is found that (i) with relatively small sample sizes, the predictions made with unweighted sampling tend to underestimate the population totals, while the ones made with weighted sampling tend to overestimate the population totals; (ii) with large sample sizes, the predictions made with weighted sampling tend to underestimate the population totals as well; (iii) with relatively small sample sizes, unweighted sampling leads to overestimating population counts within less populous areas and underestimating population counts within populous and the most populous areas; (iv) with large sample sizes, unweighted sampling leads to underestimating population counts within the least populous and populous areas but overestimating population counts within the most populous areas; (v) with relatively small sample sizes, weighted sampling re-

Table 4.2: Predicted population counts on the original scale per 15 arc-seconds gridded cell and their totals derived on the basis of the estimations made with each combination of sample sizes, sampling methods and types of priors enumerated.

| Predicted | Min | 25% | 50% | Mean | 75% | Max | Total |
|---|---|---|---|---|---|---|---|
| Size | unweighted sampling + normal priors | | | | | | |
| 1% | 0.0995 | 9.6161 | 27.9802 | 75.4609 | 99.5234 | 728.8134 | 807734 |
| 2% | 0.0608 | 7.8084 | 25.3979 | 70.1761 | 81.3847 | 1010.1440 | 751165.4 |
| 5% | 0.000 | 4.087 | 21.122 | 93.673 | 96.882 | 1134.463 | 1002677 |
| 10% | 0.000 | 4.125 | 19.918 | 107.944 | 89.138 | 1638.537 | 1155433 |
| 20% | 0.000 | 4.181 | 19.964 | 120.857 | 95.091 | 1739.967 | 1293657 |
| 50% | 0.000 | 3.993 | 18.520 | 134.359 | 107.624 | 3428.496 | 1438180 |
| Size | unweighted sampling + PC priors | | | | | | |
| 1% | 0.000 | 9.402 | 29.758 | 86.076 | 108.769 | 6867.500 | 921360.2 |
| 2% | 0.7123 | 7.9236 | 20.5141 | 58.9683 | 68.2373 | 507.7850 | 631196.5 |
| 5% | 0.0228 | 4.7492 | 18.9142 | 77.4187 | 81.0333 | 944.0085 | 828689.3 |
| 10% | 0.000 | 4.579 | 18.554 | 90.042 | 79.296 | 1145.479 | 963812.7 |
| 20% | 0.000 | 4.422 | 19.453 | 110.629 | 90.515 | 1492.890 | 1184169 |
| 50% | 0.000 | 4.057 | 18.482 | 131.392 | 105.768 | 3225.403 | 1406423 |
| Size | unweighted sampling + vague priors | | | | | | |
| 1% | 1.326 | 9.924 | 25.267 | 75.136 | 101.134 | 1060.737 | 804253.7 |
| 2% | 0.5809 | 7.6878 | 21.0950 | 62.3893 | 71.4285 | 555.7964 | 667815.4 |
| 5% | 0.00 | 4.17 | 18.63 | 84.70 | 89.10 | 1015.44 | 906624.3 |
| 10% | 0.000 | 4.292 | 18.177 | 89.485 | 84.692 | 1623.270 | 957849 |
| 20% | 0.000 | 4.371 | 19.423 | 112.005 | 91.376 | 1491.670 | 1198902 |
| 50% | 0.00 | 4.04 | 18.49 | 131.94 | 106.22 | 3230.36 | 1412303 |
| Size | weighted sampling + normal priors | | | | | | |
| 1% | 6.029 | 31.757 | 92.444 | 214.151 | 316.210 | 1167.151 | 2292270 |
| 2% | 11.43 | 47.27 | 121.30 | 243.80 | 364.96 | 1192.04 | 2609587 |
| 5% | 2.697 | 46.606 | 105.975 | 208.032 | 286.027 | 1193.520 | 2226779 |
| 10% | 2.515 | 43.290 | 93.552 | 191.271 | 233.000 | 1361.226 | 2047363 |
| 20% | 0.395 | 26.969 | 67.416 | 172.131 | 192.721 | 1646.395 | 1842487 |
| 50% | 0.00 | 12.42 | 38.12 | 151.00 | 144.10 | 2237.77 | 1616252 |
| Size | weighted sampling + PC priors | | | | | | |
| 1% | 0.00 | 32.85 | 109.61 | 232.08 | 331.96 | 1919.60 | 2484148 |
| 2% | 0.00 | 50.87 | 141.74 | 258.11 | 366.58 | 2777.84 | 2762776 |
| 5% | 0.00 | 40.41 | 107.15 | 220.79 | 273.34 | 6806.83 | 2363305 |
| 10% | 0.00 | 35.59 | 86.26 | 204.45 | 253.15 | 4369.21 | 2188442 |
| 20% | 0.00 | 21.34 | 61.82 | 183.79 | 214.14 | 8137.67 | 1967341 |
| 50% | 0.00 | 12.38 | 38.30 | 151.74 | 145.02 | 2236.38 | 1624174 |
| Size | weighted sampling + vague priors | | | | | | |
| 1% | 0.00 | 19.61 | 93.38 | 228.72 | 343.53 | 1948.86 | 2448232 |
| 2% | 0.00 | 63.52 | 151.42 | 261.37 | 356.70 | 2889.60 | 2797709 |
| 5% | 0.00 | 48.41 | 112.91 | 224.80 | 270.39 | 8229.83 | 2406306 |
| 10% | 0.00 | 42.91 | 91.88 | 208.48 | 250.50 | 6283.40 | 2231562 |
| 20% | 0.00 | 23.68 | 63.50 | 184.94 | 214.74 | 8782.50 | 1979541 |
| 50% | 0.00 | 12.50 | 38.41 | 151.51 | 144.80 | 2241.35 | 1621758 |

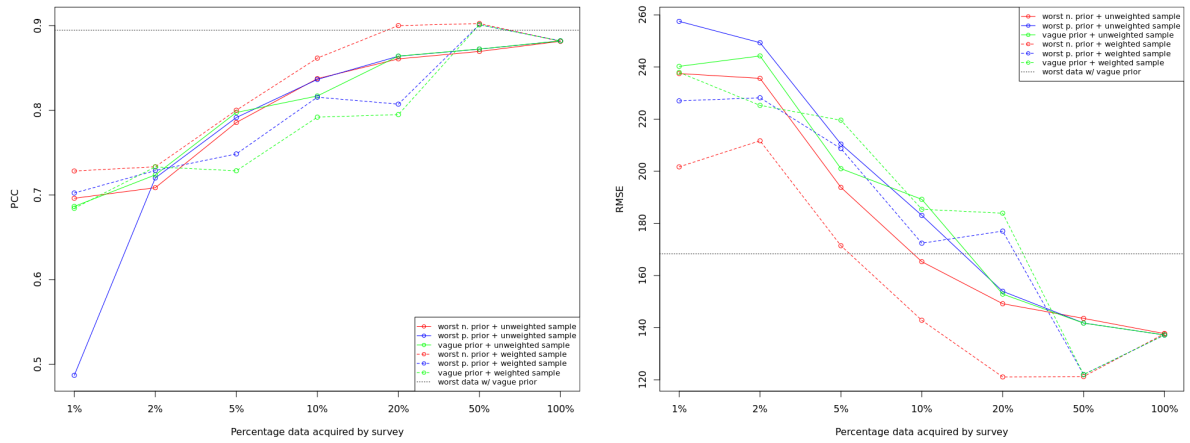
sults in overestimating population counts within the whole region; (vi) with large sample sizes, weighted sampling results in overestimating population counts within less populous areas, underestimating population counts within populous areas but overestimating population counts within the most populous areas. This is probably because (i) the DMSP-like NTL data suffer from the "overglow" effect which smooths light intensity through diffusing the brightness from the bright areas to the dark areas so cannot be regarded as a good indicator of population presence in less populous areas; (ii) the spatial random effects follow Gaussian distribution, but population counts are very likely saturated in the most populous areas. It is also found that (i) normal priors indeed help the estimations with relatively small sample sizes improve the underestimation and overestimation of population counts caused by unweighted sampling and weighted sampling respectively within the most populous areas; (ii) weighted sampling indeed reduces posterior SD within more populous areas, but meanwhile increases posterior SD within less populous areas; (iii) without utilising external information, vague priors do not reduce posterior SD around the vertices of the triangular mesh.

Figure A.14 and Figure A.15 available in Appendix A present the mapped posterior means and SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of the estimations made with "bottom-up" sample of 100% full real population data and the ones derived by projecting existing "top-down" population estimates. Table 4.3 summarises the predicted population counts on the original scale per 15 arc-seconds gridded cell and their totals derived on the basis of estimations made with "bottom-up" sample of 100% full real population data as the likelihood (i.e., the previous "best data with worst prior" case) and the ones derived by projecting existing "top-down" population estimates available originally at a 30 arc-seconds resolution for free to a finer 15 arc-seconds resolution. The

Table 4.3: Predicted population counts on the original scale per 15 arc-seconds gridded cell and their totals derived on the basis of the estimations made with "bottom-up" sample of 100% real population data and the ones derived by projecting existing "top-down" population estimates.

| Predicted | Min | 25% | 50% | Mean | 75% | Max | Total |
|---|---|---|---|---|---|---|---|
| Prior | "best data with worst prior" case | | | | | | |
| normal | 0.000 | 3.765 | 18.750 | 140.236 | 113.058 | 3039.088 | 1501086 |
| PC | 0.000 | 3.772 | 18.708 | 139.442 | 112.862 | 2974.369 | 1492586 |
| vague | 0.000 | 3.772 | 18.692 | 139.582 | 113.193 | 2971.693 | 1494089 |
| Prior | projecting "top-down" population estimates | | | | | | |
| vague | 0.000 | 6.547 | 21.267 | 93.808 | 90.129 | 1084.278 | 1004117 |

results from testing the most important idea of assessing the border between "top-down" and "bottom-up" approaches are shown as Figure 4.15, which are based on comparing the predictive performances of the "bottom-up" models (with or without existing "top-down" estimates as prior information) and the model that projects existing "top-down" estimates to a target spatial resolution as required with selection criteria PCC and RMSE. Without the estimation errors contained in the original gridded population data (i.e., the GHS-POP data at a 30 arc-seconds resolution in this case) considered, it is found that (i) in terms of PCC, the best performing "bottom-up" model, the one based on normal priors and weighted sampling, starts to slightly outperform the projected "top-down" estimates with a sample size of 20% real




(a) (b)

Figure 4.15: Predictive performances of the "bottom-up" models with increasing sample sizes and the model projecting "top-down" estimates, assessed with (a) PCC and (b) RMSE (the "worst priors" and "worst data" referred to the prior information and likelihood derived with the "top-down" population estimates available at a coarser resolution for free).

population data; (ii) an increasing sample size over 20% real population data cannot bring an apparently better result and that means a "bottom-up" "microcensus" household survey is meaningless; (iii) in terms of RMSE, the best performing "bottom-up" model, the same one based on normal priors and weighted sampling, almost outperforms the projected "top-down" estimates with a sample size of 5% real population data; (iv) only increasing the sample size to 20% real population data could the "bottom-up" model perform apparently better than the project "top-down" estimates, and that means a meaningful "bottom-up" "microcensus" household survey would cost between $10818.192 and $54090.96 in total for the provincie of Utrecht, according to the per capita cost suggested by Wardrop et al. (2018) and previously mentioned in the research design section of this chapter. The stakeholder has then to decide whether to invest such money or just to enjoy the "free lunch" (i.e., good predictive performances of the projected "top-down" estimates achieved for free).

# Chapter 5

# A Practical Test

This chapter aims to test the idea of assessing the border between "top-down" and "bottom-up" approaches and to test how different types of priors function in realising the idea of using existing "top-down" estimates as prior information. As (i) the official CBS demographic data at a 15 arc-seconds resolution were used as unknown real population counts instead of the GHS-POP data at a 15 arc-seconds resolution tested in Chapter 4; (ii) the GHS-POP data at a 30 arc-seconds resolution were again employed as existing "top-down" gridded population estimates to be projected but regarded as less precise estimates on the real population counts, this chapter is actually a test with less precise "top-down" estimates in hand.

## 5.1 Research Design

In this chapter, the core ideas, the idea of assessing the border between "top-down" and "bottom-up" approaches and the idea of combining the "top-down" and "bottom-up" approaches, remain unchanged, while the only difference from Chapter 4 is that the real population data defined in Chapter 4 (i.e., the GHS-POP data at a resolution of 15 arc-seconds) were replaced by the real population data defined in this chapter (i.e., the official georeferenced CBS population counts at a resolution of 15 arc-seconds) which are mentioned in the large-scale gridded population data section of Chapter 2. In this chapter, the real population data directly reflect true information about population distributions and counts, while the population data including but not limited to the ones shown in Table 1.1 and the ones defined as the real population data in Chapter 4 are just estimates on the real population numbers. That means the tests to be conducted in this chapter do consider the measurement errors and statistical errors contained in the existing "top-down" population estimates (i.e., the large-scale gridded population data products based on "top-down" methods are not precise estimators on the real population distributions and counts). Similarly, it was assumed the real population data could be collected in a reliable and cost-effective "microcensus" through limited household surveys in randomly selected areas. Because the weighted random sampling was also weighted by the 30 arc-seconds GHS-POP data which provide information for free, the final collected samples are just the same as the ones collected in Chapter 4 (see Figure 4.4). The out-of-sample predictive performance of each model was evaluated through PCC and RMSE against 15 arc-seconds real population data to inform qualities of the projected population estimates.

The test with less precise "top-down" population estimates in hand, presented in this chapter, is summarised as Figure 4.1.

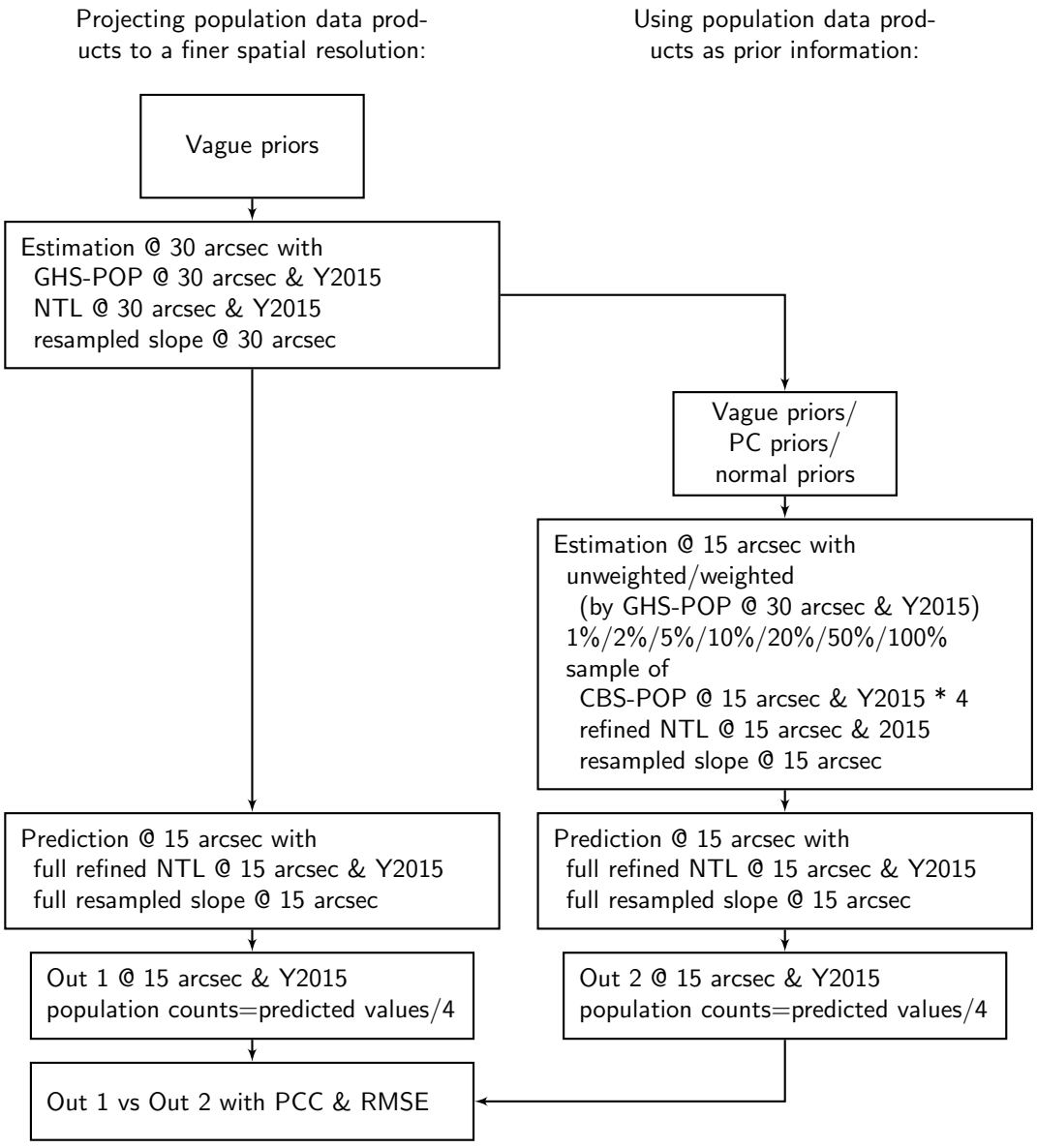



Figure 5.1: A test with less precise "top-down" population estimates in hand.

## 5.2 Exploratory Data Analysis

The GHS-POP data at resolutions of 15 arc-seconds (i.e., the real population data in Chapter 4) and the CBS population data at the same resolution (i.e., the real population data in this chapter) are visualised as Figure B.1 and available in Appendix B, with the transformation function $g_3(y_i)$ applied. It can be found that the population counts derived by the CBS are more concentrated in the inhabitable areas rather than the commercial zones (e.g., suburban shopping centres), according to the Google Street View. A comparison of summary statistics between the real population data defined in Chapter 4 and the real population data defined in this chapter is presented as Table 5.1. It is interesting to note that the population total

Table 5.1: Summary statistics of the real population data defined in Chapter 4 and the real population data defined in this chapter.

| Type | Min | 25% | 50% | Mean | 75% | Max | Total |
|---|---|---|---|---|---|---|---|
| Chapter 4 | 0.000 | 2.024 | 35.051 | 168.445 | 178.576 | 1278.376 | 1803032 |
| Chapter 5 | 0.0 | 0.0 | 10.0 | 218.4 | 70.0 | 4635.0 | 2338005 |

calculated with the real population data defined in this chapter is much larger than the one calculated with the real population data defined in Chapter 4. The PCC and RMSE between these two types of real population data are found to be 0.741099 and 365.2927 respectively, while the de jure population total within the provincie of Utrecht in 2015 provided by the CBS StatLine database is 1263572 (CBS, 2023). It implies that (i) the GHS-POP data at a resolution of 15 arc-seconds, considered as the real population data in Chapter 4, might not be precise estimates on the real population distributions and counts; (ii) the official CBS population data at a resolution of 15 arc-seconds, considered as the real population data in this chapter, might not be very real (certainly they were not, due to the artificial ambiguity as mentioned in the large-scale gridded population data in Chapter 2).

A plot of the classic and robust versions of empirical semivariogram and directional empirical semivariogram values for the CBS population data at a 15 arc-seconds resolution on the transformed scale $g_3(y_i)$ is also shown in Figure 4.2.

## 5.3 Counterpart: Yet Another Hierarchical Model

An important motivation for using the CBS population data as the real population data is to compare the model proposed in this master's thesis with the Bayesian hierarchical model proposed by Leasure et al. (2020), because the CBS demographic data contain the information of school density and average household size per cell required by Leasure et al. (2020) as the geospatial ancillary covariates available nationwide.

Previously mentioned in the research design section of Chapter 4, the "microcensus" clusters with slightly different areal sizes used in real practices were simplified as the "microcensus" gridded cells with the same areal size, so the drawn samples could be understood as either population densities or population counts. Again, very similar to the LGCP model, the average population counts per unit of space (i.e., population density) could be modelled with an intensity function $\lambda_i$, presented as Equation 5.1 and Equation 5.2, but the spatial effects are

no longer considered by the hierarchical model proposed by Leasure et al. (2020).

$$y_i \sim \text{Pois}(\lambda_i) \tag{5.1}$$

$$\ln(\lambda_i) = \alpha_{g(t)} + \sum_{k=1}^{K} \beta_k x_{ki} \tag{5.2}$$

where $\alpha_{g(t)} \sim \mathcal{N}\left(\mu_{g(t)}, \sigma^2_{g(t)}\right)$ with $\sigma^2_{g(t)}$ quantifying random variations in population counts that are not explained by the covariates, and solving the dispersion problem likely associated with the Poisson distribution in Equation 5.1.

Leasure et al. (2020) treated $\mu_{g(t)}$ as average population counts for settlement type $t$ in each gemeente $g$, drawn from the distribution of average population counts for that settlement type throughout the provincie of Utrecht, shown as Equation 5.3, and treated $\mu_t$ as average population counts for each settlement type $t$ drawn from the distribution of population counts among all the "microcensus" gridded cells, shown as Equation 5.4 and Equation 5.5.

$$\mu_{g(t)} \sim \mathcal{N}(\mu_t, \theta_t) \tag{5.3}$$

$$\mu_t \sim \mathcal{N}(\mu, \theta) \tag{5.4}$$

$$\theta_t \sim \mathcal{U}(0, \theta) \tag{5.5}$$

Residual variation $\sigma_{g(t)}$ for each gemeente $g$ would be treated with a similar hierarchical structure, presented as Equation 5.6 - Equation 5.8.

$$\sigma_{g(t)} \sim \text{Half-Normal}(\sigma_t, \epsilon_t) \tag{5.6}$$

$$\sigma_t \sim \mathcal{N}(\sigma, \epsilon) \tag{5.7}$$

$$\epsilon_t \sim \mathcal{U}(0, \epsilon) \tag{5.8}$$

The hyperparameters estimated are $\mu$, $\theta$, $\sigma$ and $\epsilon$, while minimally informative priors defined by Leasure et al. (2020) are $\beta_k \sim \mathcal{N}(0, 5)$, $\mu \sim \mathcal{N}(0, 31.6)$, $\sigma \sim$ Half-Normal$(0, 31.6)$, $\theta \sim \mathcal{U}(0, 1000)$ and $\epsilon \sim \mathcal{U}(0, 1000)$. The hierarchical random intercept $\alpha_{g(t)}$ actually functions like the spatial random effect proposed in this master's thesis, which also accounts for spatial autocorrelation inherent in data from nearby clusters.

One of the major differences between the hierarchical model to be tested in this chapter and the one proposed by Leasure et al. (2020) is that the random intercept used in this chapter is hierarchical only at two levels rather than four levels (i.e., settlement type, region, state and local government area), because the provincie of Utrecht is much smaller than Nigeria. The differences between the covariates used in this chapter for the simplified hierarchical model and the ones used by Leasure et al. (2020) include (i) the GHS-POP data at a resolution of 30 arc-seconds are considered as the large-scale population data product available in this chapter instead of the WorldPop population data at a resolution of 3 arc-seconds; (ii) school densities calculated with the map of schools in Nigeria provided eHealth Africa are replaced by average numbers of primary schools within 1 km by road for all residents of an area provided by the CBS georeferenced demographic data; (iii) household sizes interpolated with Demographic Health Survey results are substituted with numbers of inhabitants living in private households divided by the numbers of private households given by the CBS georeferenced demographic data; (iv) settlement types derived by applying feature extraction to WorldView 2, Pléiades 1A and Pléiades 1B satellite imagery pansharpened at a 0.5 m spatial resolution are replaced

by settlement typologies delineated, classified into multiple types (from urban centre grid cell to very low density rural grid cell) via a logic on the basis of population clusters and built-up densities, and provided as GHS Settlement Model Layers (GHS-SMOD) R2022A dataset (Schiavina et al., 2022) available at a 30 arc-seconds resolution for 2015 and for free; (v) settled, residential and nonresidential settled areas within a 1 km radius derived with the same feature extraction technologies are substituted with built-up fraction estimates, produced from 10 m resolution Sentinel-2 image composite and symbolic ML and provided as GHS Built-up Surface Grid (GHS-BUILT-S) R2022A dataset (Pesaresi and Politis, 2022) available at a 3 arc-seconds resolution for 2015 and for free.

The hierarchical models discussed in this section were estimated with MCMC methods in Just Another Gibbs Sampler (JAGS) through the `R-runjags` package. Three MCMC chains, 1000 iterations in the adaptive phase for optimising samplers, 1000 iterations thrown away (burn-in) at the beginning of each MCMC run and 2000 iterations in the sampling phase were specified. Convergences are checked with trace plots.

The underdispersion problem suffered by the LGCP models is presented in the results section of this chapter with the real population data as well.

## 5.4 Results

The triangular mesh used for all the calculations is visualised as Figure 4.3. Attention is paid to the "best data with worst prior" case and the "worst data with best prior" case, as this time the new real population data (i.e., the 15 arc-seconds CBS population data) were treated as the likelihood for the former case and the prior information for the later case. The posterior distributions of the hyperparameters (i.e., $\sigma_e^2$, $\sigma^2$ and $r$) and parameters (i.e., $\beta_0$, $\beta_1$ and $\beta_2$) derived with the estimations are shown as Figure 5.2. The same conclusions could be drawn: (i) more informative priors result in the higher "peaks" of posterior distributions; (ii) PC priors and normal priors reduce variances of posterior distributions; (iii) no conclusion could be made on which case is better, because amount of information contained in the real population data is also four times more than that in the GHS-POP data at a 30 arc-seconds resolution.

Following the procedures as introduced in the research design section of Chapter 4 and summarised as Figure 5.1, thirty-six estimations were made with each combination of six sample sizes, two sampling methods and three types of priors enumerated, to assess the idea of using "top-down" population data products as prior information. In this chapter, the estimations were made with the sampled CBS population data instead of the sampled GHS-POP data. The posterior distributions of the hyperparameters (i.e., $\sigma_e^2$, $\sigma^2$ and $r$) and parameters derived with the estimations on the basis of (i) unweighted sampling and normal priors, (ii) unweighted sampling and PC priors, (iii) unweighted sampling and vague priors, (iv) weighted sampling and normal priors, (v) weighted sampling and PC priors, and (vi) weighted sampling and vague priors are shown as Figure 5.3, Figure 5.4, Figure 5.5, Figure 5.6, Figure 5.7 and Figure 5.8. It is found that, although the measurement errors and statistical errors contained in the existing "top-down" population estimates (i.e., the 30 arc-seconds GHS-POP data used as the prior information) are considered in this chapter, normal priors still perform much better than PC priors which are merely a slightly better choice than vague priors.

Table 5.2 summarises the predicted population counts on the original scale per 15 arc-seconds gridded cell and their totals derived on the basis of the thirty-six estimations, comparable to

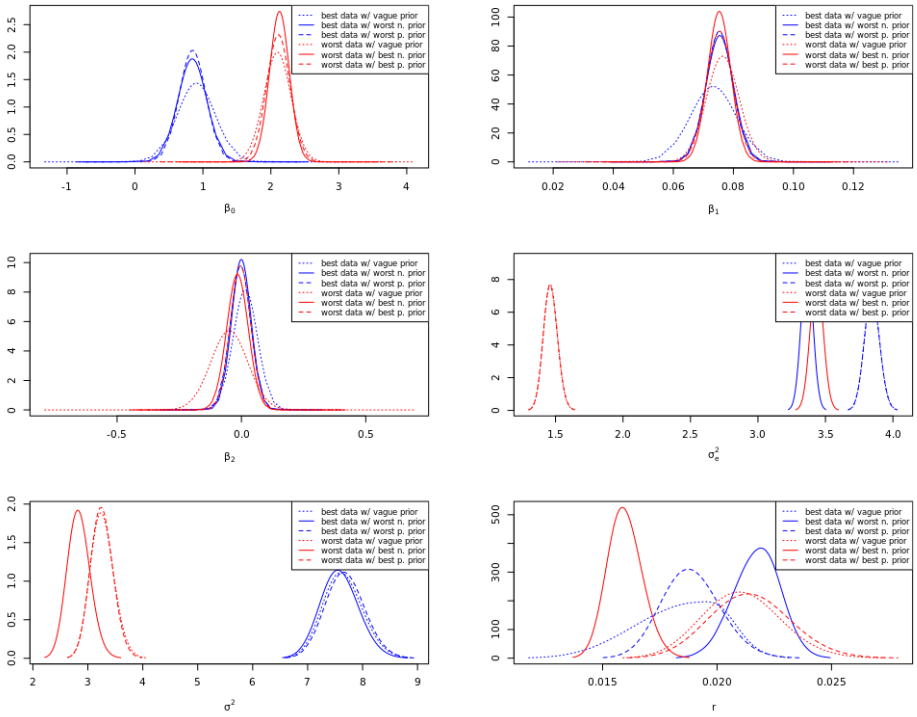



Figure 5.2: Posterior distributions of parameters and hyperparameters for estimations made with the "best data with worst prior" case in blue and the "worst data with best prior" case in red.

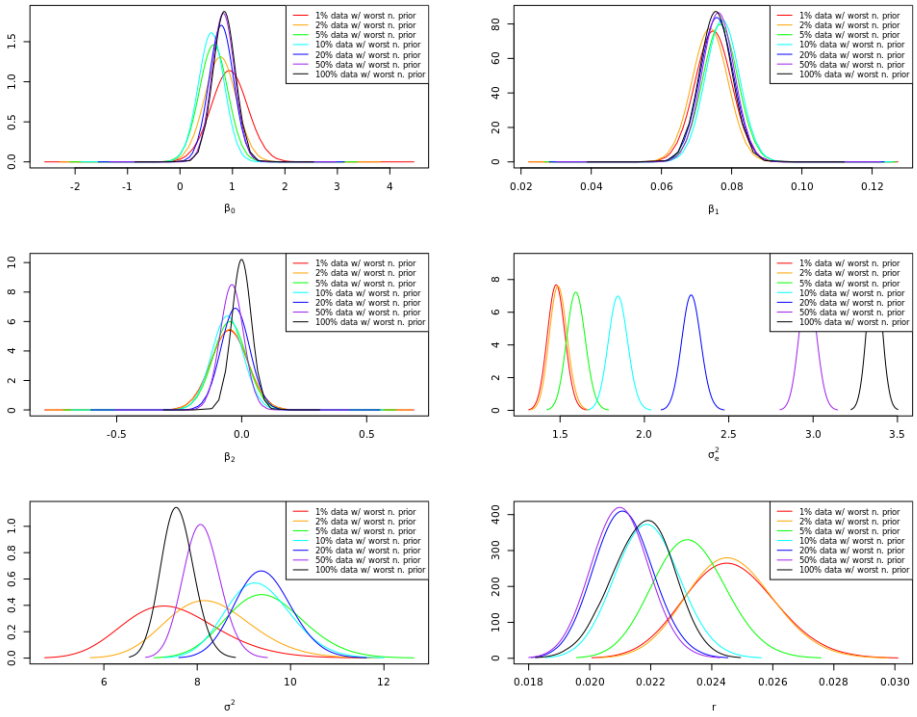



Figure 5.3: Posterior distributions of parameters and hyperparameters for estimations made with unweighted samples of 1%, 2%, 5%, 10%, 20%, 50% and 100% real population data and normal priors in red, orange, green, cyan, blue, purple and black.

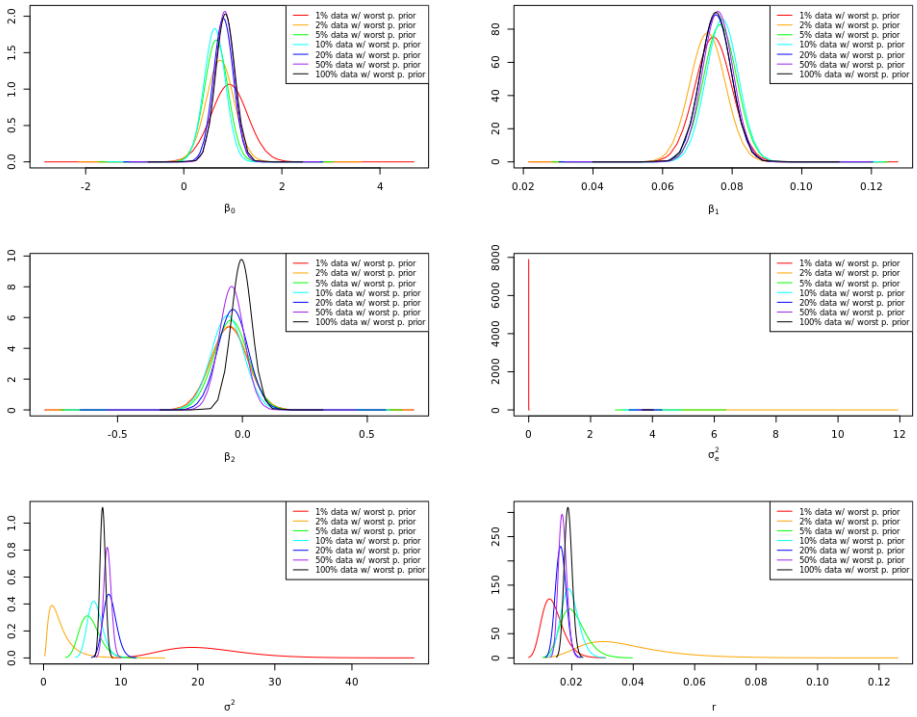



Figure 5.4: Posterior distributions of parameters and hyperparameters for estimations made with unweighted samples of 1%, 2%, 5%, 10%, 20%, 50% and 100% real population data and PC priors in red, orange, green, cyan, blue, purple and black.

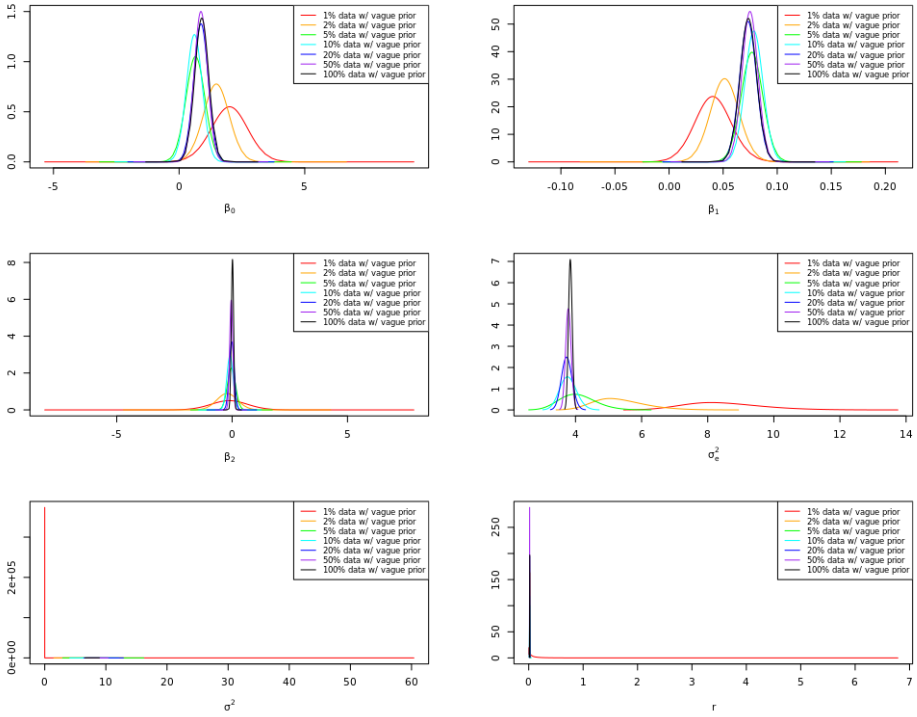



Figure 5.5: Posterior distributions of parameters and hyperparameters for estimations made with unweighted samples of 1%, 2%, 5%, 10%, 20%, 50% and 100% real population data and vague priors in red, orange, green, cyan, blue, purple and black.

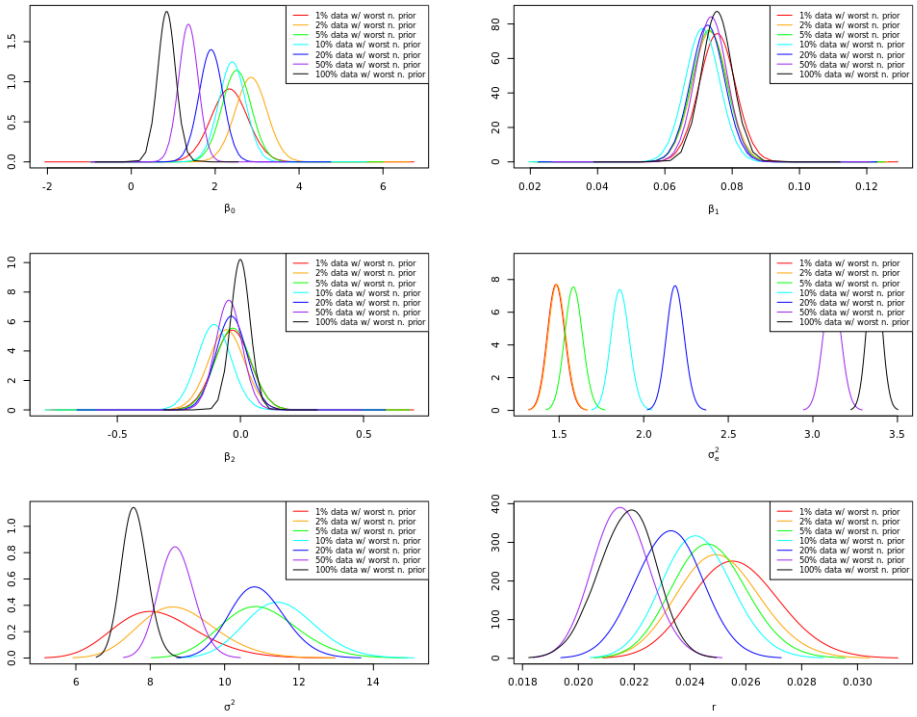



Figure 5.6: Posterior distributions of parameters and hyperparameters for estimations made with weighted samples of 1%, 2%, 5%, 10%, 20%, 50% and 100% real population data and normal priors in red, orange, green, cyan, blue, purple and black.

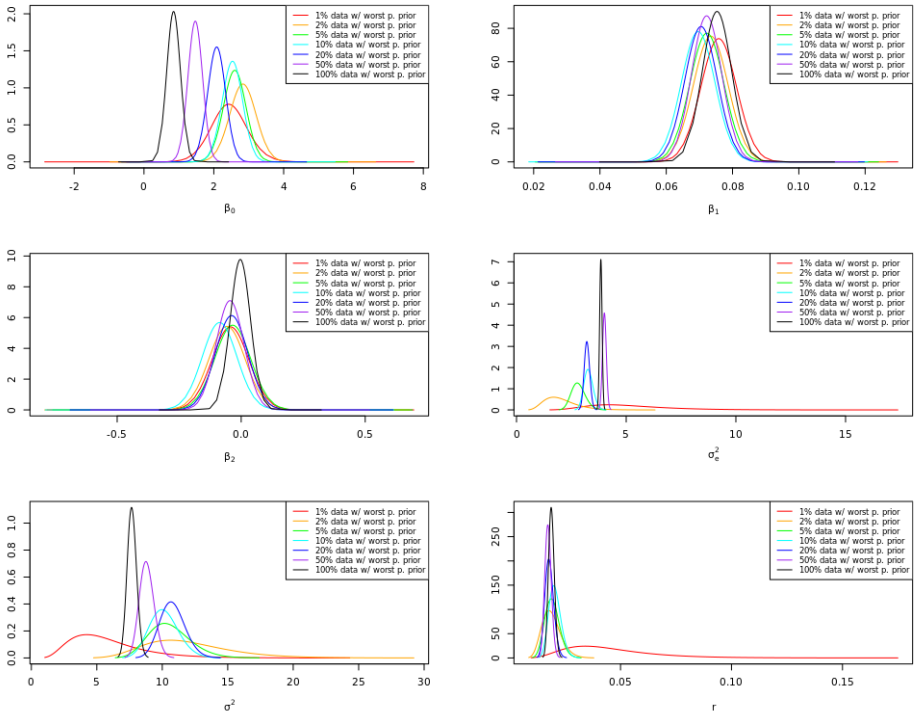



Figure 5.7: Posterior distributions of parameters and hyperparameters for estimations made with weighted samples of 1%, 2%, 5%, 10%, 20%, 50% and 100% real population data and PC priors in red, orange, green, cyan, blue, purple and black.

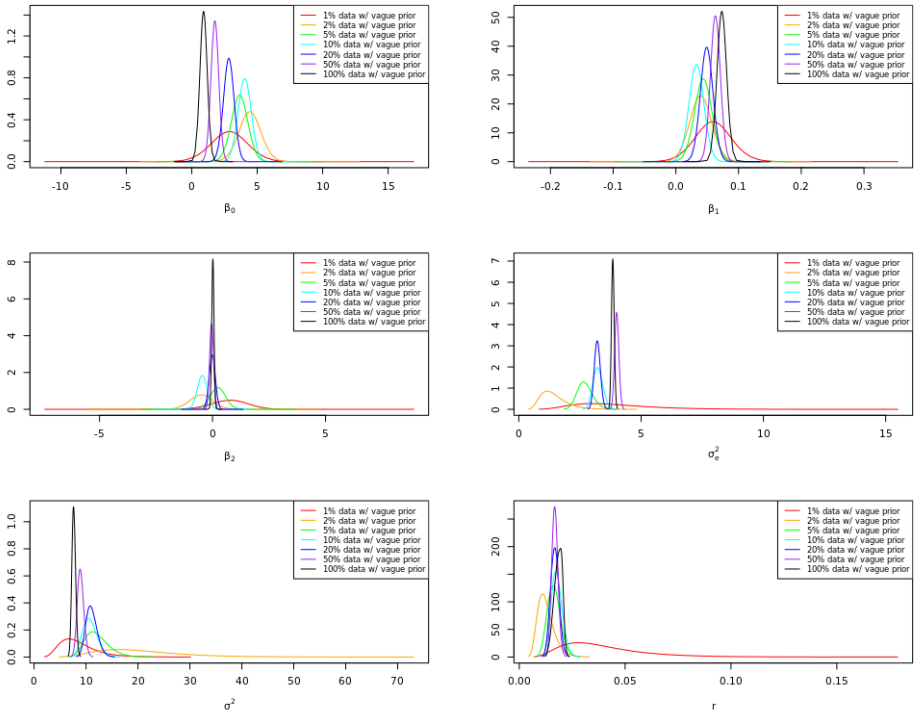



Figure 5.8: Posterior distributions of parameters and hyperparameters for estimations made with weighted samples of 1%, 2%, 5%, 10%, 20%, 50% and 100% real population data and vague priors in red, orange, green, cyan, blue, purple and black.

the data in Table 5.1. Figure B.2 and Figure B.3, Figure B.4 and Figure B.5, Figure B.6 and Figure B.7, Figure B.8 and Figure B.9, Figure B.10 and Figure B.5, and Figure B.12 and Figure B.13 available in Appendix B show the mapped posterior means and SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of the above-mentioned thirty-six estimations made with (i) unweighted sampling and normal priors, (ii) unweighted sampling and PC priors, (iii) unweighted sampling and vague priors, (iv) weighted sampling and normal priors, (v) weighted sampling and PC priors, and (vi) weighted sampling and vague priors, and different sample sizes, comparable to the map of the real population data defined in this chapter and shown as Figure B.1b. It is found that (i) with relatively small "bottom-up" sample sizes, the predictions made with unweighted sampling tend to severely underestimate the population totals, while the ones made with weighted sampling tend to underestimate the population totals as well but more mildly; (ii) increasing relatively small sample sizes to large sample sizes results in larger predicted population totals with unweighted samples and smaller predicted population totals with weighted samples; (iii) with relatively small sample sizes, unweighted sampling brings about overestimating population counts within less populous areas and underestimating population counts within populous and the most populous areas; (iv) with large sample sizes, unweighted sampling brings about overestimating population counts within less populous and the most populous areas but underestimating population counts within populous areas; (v) with relative small sample sizes, weighted sampling results in overestimating population counts within the whole region; (vi) with large sample sizes, weighted sampling results in overestimating population counts within less populous areas, mildly overestimating or underestimating population counts

Table 5.2: Predicted population counts on the original scale per 15 arc-seconds gridded cell and their totals derived on the basis of the estimations made with each combination of sample sizes, sampling methods and types of priors enumerated.

| Predicted | Min | 25% | 50% | Mean | 75% | Max | Total |
|---|---|---|---|---|---|---|---|
| Size | unweighted sampling + normal priors | | | | | | |
| 1% | 0.000 | 1.873 | 6.112 | 22.684 | 21.553 | 1590.337 | 242809.2 |
| 2% | 0.000 | 1.276 | 5.036 | 27.660 | 17.360 | 2952.978 | 296070.2 |
| 5% | 0.000 | 0.707 | 4.219 | 61.541 | 23.706 | 3645.940 | 658732.1 |
| 10% | 0.000 | 0.573 | 3.525 | 90.238 | 23.923 | 5246.624 | 965909.1 |
| 20% | 0.000 | 0.732 | 4.178 | 122.525 | 26.515 | 11032.710 | 1311508 |
| 50% | 0.000 | 0.778 | 3.724 | 145.519 | 27.146 | 11709.057 | 1557631 |
| Size | unweighted sampling + PC priors | | | | | | |
| 1% | 0.000 | 1.753 | 6.057 | 37.179 | 21.649 | 15121.926 | 397962 |
| 2% | 0.1491 | 1.6444 | 4.4312 | 10.5493 | 13.8139 | 119.7585 | 112920.1 |
| 5% | 0.000 | 1.194 | 4.207 | 27.393 | 16.652 | 1105.840 | 293212.1 |
| 10% | 0.0000 | 0.9496 | 3.7298 | 47.6789 | 17.8803 | 2250.0668 | 510355.1 |
| 20% | 0.000 | 0.944 | 4.110 | 81.327 | 21.801 | 4491.000 | 870523.9 |
| 50% | 0.000 | 0.835 | 3.753 | 129.899 | 25.916 | 9499.796 | 1390443 |
| Size | unweighted sampling + vague priors | | | | | | |
| 1% | 0.4192 | 3.2333 | 5.7505 | 8.0426 | 11.8706 | 25.8005 | 86087.59 |
| 2% | 0.000 | 2.047 | 4.622 | 9.175 | 11.147 | 152.047 | 98211.91 |
| 5% | 0.000 | 1.173 | 4.217 | 28.226 | 17.074 | 1158.420 | 302132.6 |
| 10% | 0.0000 | 0.9327 | 3.7475 | 48.1127 | 18.1828 | 2254.0922 | 514998.5 |
| 20% | 0.000 | 0.942 | 4.119 | 81.157 | 21.846 | 4533.860 | 868699.4 |
| 50% | 0.000 | 0.837 | 3.755 | 129.969 | 25.928 | 9525.143 | 1391192 |
| Size | weighted sampling + normal priors | | | | | | |
| 1% | 0.000 | 7.863 | 20.886 | 124.248 | 81.816 | 9371.154 | 1329948 |
| 2% | 0.053 | 12.405 | 34.436 | 174.679 | 129.137 | 19876.414 | 1869759 |
| 5% | 0.000 | 7.374 | 25.419 | 195.731 | 117.391 | 10909.539 | 2095100 |
| 10% | 0.000 | 6.645 | 22.465 | 190.086 | 108.706 | 11180.022 | 2034680 |
| 20% | 0.000 | 3.308 | 13.621 | 177.795 | 81.093 | 9298.961 | 1903115 |
| 50% | 0.000 | 1.835 | 7.456 | 158.886 | 43.748 | 8629.881 | 1700719 |
| Size | weighted sampling + PC priors | | | | | | |
| 1% | 0.7081 | 9.9442 | 25.4356 | 92.6596 | 81.6633 | 2050.1587 | 991828.4 |
| 2% | 0.195 | 12.619 | 33.340 | 156.118 | 121.050 | 12927.428 | 1671086 |
| 5% | 0.000 | 9.139 | 24.585 | 157.734 | 101.105 | 7993.459 | 1688387 |
| 10% | 0.000 | 8.055 | 21.648 | 157.182 | 90.227 | 8841.332 | 1682476 |
| 20% | 0.000 | 4.349 | 13.643 | 161.899 | 70.643 | 7176.802 | 1732962 |
| 50% | 0.000 | 2.009 | 7.534 | 150.191 | 42.559 | 7668.436 | 1607646 |
| Size | weighted sampling + vague priors | | | | | | |
| 1% | 0.44 | 13.75 | 31.62 | 149.16 | 97.89 | 177583.28 | 1596580 |
| 2% | 0.00 | 29.01 | 51.19 | 176.56 | 120.19 | 25473.16 | 1889927 |
| 5% | 0.00 | 16.32 | 33.06 | 160.58 | 104.28 | 8204.73 | 1718894 |
| 10% | 0.00 | 12.94 | 28.55 | 159.00 | 91.65 | 9194.88 | 1701896 |
| 20% | 0.000 | 5.307 | 14.937 | 161.698 | 71.825 | 7270.640 | 1730814 |
| 50% | 0.000 | 2.064 | 7.553 | 149.787 | 42.561 | 7684.637 | 1603323 |

within populous areas, and overestimating population counts within the most populous areas. It is probably because (i) the fact that the CBS recorded population counts only in the inhabitable areas rather than all the built-up areas where the GHSL programme allocated census data makes unweighted samples and weighted samples with large sample sizes more easily be collected in uninhabitable areas; (ii) spatial discontinuity between population-more-concentrated habitable areas and uninhabitable areas is harder to model with Matérn parameters than that between population-less-concentrated built-up areas and undeveloped areas (i.e., the GHSL programme actually smoothed out census data by assigning population counts into built-up but uninhabitable areas); (iii) due to an unknown reason, the population total provided by CBS outnumbers the one based on the Dutch census, which means the real population data are not extremely reliable. It is also fount that (i) normal priors indeed help the estimations with relatively small sample sizes stabilise the predictions made within the most populous areas; (ii) posterior SD associated with locations not close to the sampled locations show larger values with the estimations based on the real population data defined in this chapter than the ones based on the real population data applied in Chapter 4.

Figure B.14 available in Appendix B shows the mapped posterior means and SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of the estimations made with "bottom-up" sample of 100% full real population data. Table 5.3 summarises the predicted population counts on the original scale per 15 arc-seconds gridded cell and their totals derived on the basis of estimations made with "bottom-up" sample of 100% full real population data as the likelihood (i.e., the previous "best data with worst prior" case) and the ones derived with the hierarchical models proposed by Leasure et al. (2020) and presented in the counterpart section of this chapter. It is found that, regarding the hierarchical models proposed by Leasure et al. (2020), (i) the predicted population counts per cell never go to zero, while the population totals predicted with relatively large sample sizes could be close to the real values; (ii) unweighted sampling and weighted sampling lead to underestimating and overestimating population counts within populous areas respectively, while both sampling methods result in severely overestimating population counts in the most populous areas. A new assessment of the LGCP model was made with the real population data, and the histograms of the PIT measure are presented as Figure 5.9 for estimations with the CBS population data at a 15 arc-seconds resolution. Still, the LGCP model suffers an underdispersion problem. Again, the results from testing the most important idea of assessing the border between "top-down" and "bottom-up" approaches are shown as Figure 5.10, which are based on comparing the predictive performances of the "bottom-up" models proposed in this master's thesis (with or without existing "top-down" estimates as prior information), the "bottom-up" models proposed by Leasure et al. (2020) and the model that projected existing "top-down" estimates to a target spatial resolution as required with selection criteria PCC and RMSE. As mentioned firstly in the research design section of Chapter 4 and later in the research design section of Chapter 5, the 30 arc-seconds GHS-POP data produced with "top-down" methods are not able to very precisely reflect the information of population distributions and counts contained in the real population data (i.e., the official georeferenced CBS population counts at a resolution of 15 arc-seconds), so there is not doubt that (i) the predictive performances of the projected population estimates drops become worse; (ii) compared with more conservative PC priors, more aggressive normal priors based on the GHS-POP data at a resolution of 30 arc-seconds lead to worse predictions, except for unweighted samples with very small sample sizes in terms of PCC, unweighted samples with relatively small sample sizes in terms of RMSE

Table 5.3: Predicted population counts on the original scale per 15 arc-seconds gridded cell and their totals derived on the basis of the estimations made with "bottom-up" sample of 100% real population data and the ones derived with the hierarchical models proposed by Leasure et al. (2020).

| Predicted | Min | 25% | 50% | Mean | 75% | Max | Total |
|---|---|---|---|---|---|---|---|
| Prior | "best data with worst prior" case | | | | | | |
| normal | 0.000 | 0.824 | 3.627 | 147.576 | 26.149 | 9130.637 | 1579651 |
| PC | 0.000 | 0.835 | 3.647 | 143.057 | 25.928 | 8493.386 | 1531284 |
| vague | 0.000 | 0.833 | 3.638 | 143.208 | 25.875 | 8567.295 | 1532899 |
| Size | Leasure et al. (2020) + unweighted sampling + vague pirors | | | | | | |
| 1% | 0.31 | 0.32 | 0.42 | 169.27 | 58.55 | 42376.97 | 1811905 |
| 2% | 0.29 | 0.30 | 0.45 | 243.04 | 52.34 | 83411.44 | 2601449 |
| 5% | 0.39 | 0.40 | 0.55 | 243.93 | 64.83 | 81085.55 | 2610980 |
| 10% | 0.33 | 0.33 | 0.45 | 207.08 | 57.33 | 68511.66 | 2216635 |
| 20% | 0.38 | 0.39 | 0.55 | 174.60 | 68.84 | 37744.26 | 1868882 |
| 50% | 0.38 | 0.39 | 0.55 | 197.43 | 65.79 | 50504.92 | 2113305 |
| Size | Leasure et al. (2020) + weighted sampling + vague pirors | | | | | | |
| 1% | 0.64 | 0.65 | 0.85 | 390.10 | 210.56 | 264823.22 | 4175616 |
| 2% | 1.29 | 1.31 | 1.68 | 312.43 | 186.77 | 66245.97 | 3344286 |
| 5% | 0.80 | 0.81 | 1.02 | 249.55 | 151.90 | 82610.84 | 2671210 |
| 10% | 0.71 | 0.72 | 0.90 | 209.52 | 130.52 | 67102.30 | 2242652 |
| 20% | 0.64 | 0.65 | 0.86 | 196.49 | 103.42 | 38889.17 | 2103178 |
| 50% | 0.79 | 0.80 | 1.07 | 207.09 | 84.56 | 46734.69 | 2216740 |
| Size | Leasure et al. (2020) + "best data" + vague pirors | | | | | | |
| 100% | 0.40 | 0.40 | 0.56 | 188.09 | 66.84 | 45363.79 | 2013322 |

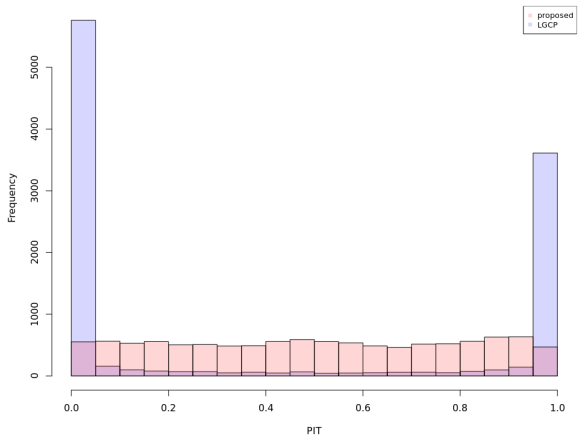
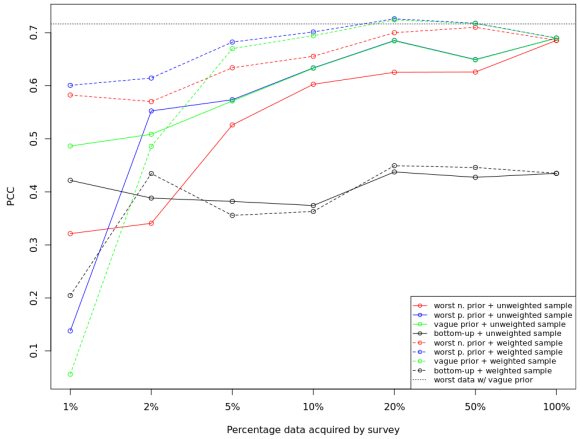
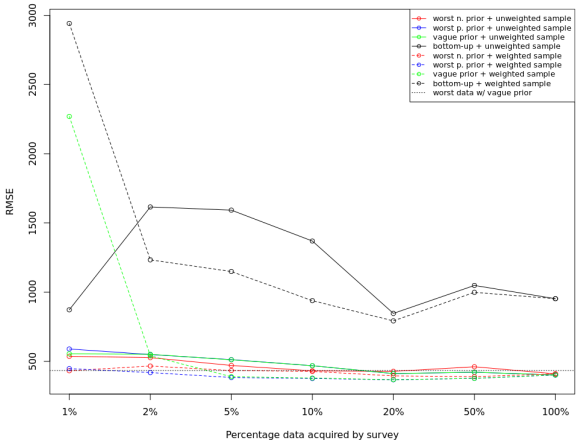


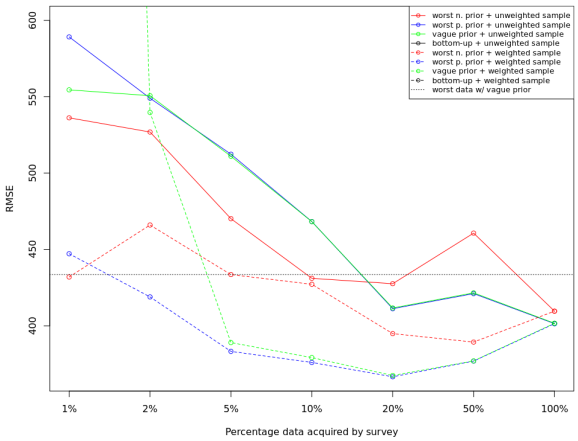

Figure 5.9: Histograms of the cross-validated PIT measure for the proposed model in red and the LGCP model in blue, derived with the CBS population data at a 15 arc-seconds resolution (i.e., the real population data defined in this chapter).




(a)




(b)




(c)

Figure 5.10: Predictive performances of the "bottom-up" models proposed in this master's thesis and the "bottom-up" models proposed by Leasure et al. (2020) with increasing sample sizes and the model projecting "top-down" estimates, assessed with (a) PCC and (b, c) RMSE with different Y-axis maximums specified (the "worst priors" and "worst data" referred to the prior information and likelihood derived with the "top-down" population estimates available at a coarser resolution for free, while the "bottom-up" referred to the "bottom-up" models proposed by Leasure et al. (2020)).

and weighted samples with very small sample sizes in terms of RMSE. With the estimation errors contained in the original gridded population data (i.e., the GHS-POP data at a 30 arc-seconds resolution in this case) considered, it is found that (i) in terms of PCC, the best performing "bottom-up" model, the one based on PC priors and weighted sampling, starts to slightly outperform the "top-down" estimates with a sample size of 20% real population data; (ii) an increasing sample size over 20% real population data cannot bring an obviously better result and that means a "bottom-up" "microcensus" household survey is meaningless; (iii) in terms of RMSE, the best performing "bottom-up" model, the same one based on PC priors and weighted sampling, outperforms the projected "top-down" estimates with a sample size between 1% and 2% real population data; (iv) increasing the sample size to 5% would enable the "bottom-up" model to perform obviously better than the "top-down" estimates, and that means a meaningful "bottom-up" "microcensus" household survey would cost between $3507.0075 and $17535.0375 in total for the provincie of Utrecht, according to the per capita cost suggested by Wardrop et al. (2018) and previously mentioned in the research design section of Chapter 4. The stakeholder has then to decide whether to invest such money or just to enjoy the "free lunch" (i.e., good predictive performances of the projected "top-down" estimates achieved for free). It has to be noted that a meaningful "bottom-up" "microcensus" household survey does not necessarily mean achieving good predictive performances, because the best predictive performances that a "bottom-up" model based on the real population data defined in Chapter 4 could reach would much better than the ones that a "bottom-up" model based on the real population data defined in this chapter could reach. How nice a "free lunch" is and how well a "bottom-up" model could perform would largely depend on how well the large-scale gridded population data products based on "top-down" methods and available for free estimate the real population distributions and counts.

Besides, it is interesting to find out that the spatial models proposed in this master's thesis perform much better than the non-spatial models with a hierarchical structure proposed by Leasure et al. (2020), only except for the estimations made with very small sample sizes and few combinations of prior types and sampling methods. The value of the methods proposed in this master's thesis could thus be confirmed.

# Chapter 6

# A Temporal Extension

This chapter aims to extend the idea of projecting exist "top-down" population estimates, firstly mentioned in the research purpose section of Chapter 1, by projecting those estimates to the temporary resolutions originally not available.

## 6.1 Spatiotemporal Modelling

When $(\boldsymbol{s}, t) \in \mathcal{D} \subset \mathbb{R}^2 \times \mathbb{R}$, the model could be extended to a dynamic regression and then Equation $(3.3)$ could be rewritten as Equation $(6.1)$

$$\eta_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2i} + \omega_{0it} + x_{1it}\omega_{1it} \tag{6.1}$$

where $\omega_{0it}$ and $\omega_{1it}$ indicate two a priori assumed independent latent spatiotemporal processes at location $s_i$ and time knot $t = 2, ..., T$ which change in time with first-order autoregression dynamics and spatially correlated innovations, with the later one weighted by local NTL strength of that time $x_{1it}$.

Then two latent processes could be presented as Equation $(6.2)$ and Equation $(6.3)$

$$\omega_{0it} = \rho_0 \omega_{0i(t-1)} + \sum_{g=1}^{G} A_{0igt}\tilde{\xi}_{0gt} \tag{6.2}$$

$$\omega_{1it} = \rho_1 \omega_{1i(t-1)} + \sum_{g=1}^{G} A_{1igt}\tilde{\xi}_{1gt} \tag{6.3}$$

where $(\rho_0, \rho_1) \in (-1, 1)$. Moreover, GFs $\xi_{0it}$ and $\xi_{1it}$ are assumed to be temporally independent and characterised by Equation $(6.4)$ and Equation $(6.5)$.

$$\mathsf{Cov}(\xi_{0it}, \xi_{0ju}) = \begin{cases} 0 & \text{if} \quad t \neq u \\ \mathsf{Cov}(\xi_{0i}, \xi_{0j}) & \text{if} \quad t = u \end{cases} \tag{6.4}$$

$$\mathsf{Cov}(\xi_{1it}, \xi_{1ju}) = \begin{cases} 0 & \text{if} \quad t \neq u \\ \mathsf{Cov}(\xi_{1i}, \xi_{1j}) & \text{if} \quad t = u \end{cases} \tag{6.5}$$

Consequently, $\boldsymbol{\omega}_{it} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\omega)$ is obtained, where $\boldsymbol{\Sigma}_\omega = \begin{pmatrix} \boldsymbol{Q}_0^{-1}(1-\rho_0^2)^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}_1^{-1}(1-\rho_1^2)^{-1} \end{pmatrix}$.

## 6.2 Research Design

As introduced in the "top-down" population mapping section of Chapter 1, projecting forward or backward from census data to a target time of interest remains a challenge due to lack of time-consistent time-specific ancillary data and use of temporally implicit or invariant ancillary data. As introduced in the refining population mapping section of Chapter 1, the time-sensitive applications (e.g., measuring substantially altered demand for social services) popularise the use of time-consistent ancillary data which hold a correlation with population and a high refresh rate (i.e., the NTL data). As mentioned in the research purpose section of Chapter 1, this chapter is dedicated to projecting existing gridded population estimates available for free to a target temporary resolution.

In this chapter, improving the temporal resolution of gridded population data from three years to one year is raised as an example, because the finest temporal resolution at which large-scale gridded population data products could so far be produced is one year, available from 2000 to 2020 (i.e., the WorldPop data; see Table 1.1). Meanwhile, the harmonised DMSP-like NTL data provided by Li et al. (2020) are only available annually from 1992 to 2018, and that means temporal matches between those two kinds of data are only available from 2000 to 2018 (i.e., 19 time knots). In a real practice, the user of the proposed approach may, for instance, downscale gridded population data available yearly to monthly, and the effectiveness of such a projection largely depends on how similar the known data at each time knot are and how well the NTL data could capture the temporal variations of the population distributions and counts.

The above-mentioned 19 time knots are divided into three groups, one estimation group and two validation groups. Population counts within 30 arc-seconds grid cells within the territory of county of Nairobi of Kenya in the years of 2001, 2004, 2007, 2010, 2013 and 2016 are used for estimation, while those in the years of 2002, 2005, 2008, 2011, 2014 and 2017 and those in the years of 2003, 2006, 2009, 2012, 2015 and 2018 are retained as Group 1 and Group 2 for validating out-of-sample predictions with PCC and RMSE. The two validation groups are proposed for assessing whether predictive performances are worse at time knots farther away from the time knots reserved for estimation, because it is interesting to know how the model would perform when projecting population counts to each sub-yearly time knot with unequal temporal distance to the census date. In this case, linear predictors derived from Equation 6.1 incorporate two white noise processes $\xi_{0it}$ and $\xi_{1it}$, which are a priori zero-mean Matérn GFs and parts of two spatiotemporal random effects $\omega_{0it}$ and $\omega_{1it}$ whose autocorrelation structures are defined by $\rho_0$ and $\rho_1$. The reason for applying this dynamic regression model, where the covariates are also modelled as time series, is to consider a possible difference in the growth rates of population counts and NTL strengths in Nairobi.

The temporal extension presented in this chapter is summarised as Figure 6.1.

## 6.3 Exploratory Data Analysis

The WorldPop data at a resolution of 30 arc-seconds for the years of 2002, 2005, 2008, 2011, 2014 and 2017 (i.e., Group 1) and the years of 2003, 2006, 2009, 2012, 2015 and 2018 (i.e., Group 2) are visualised as Figure C.1 and Figure C.2 respectively and available in Appendix C, with the transformation function $g_3(y_i)$ applied. Summary statistics of these two groups

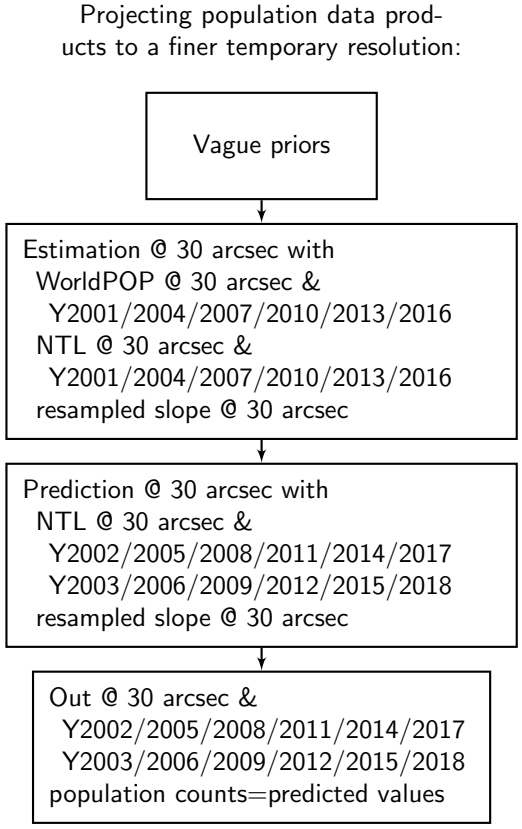



Figure 6.1: A temporal extension.

are presented as Table 6.1. The observed 16 years witness a continuous growth of population

Table 6.1: Summary statistics of the WorldPop data.

| WorldPop | Min | 25% | 50% | Mean | 75% | Max | Total |
|---|---|---|---|---|---|---|---|
| Year | | | | Group 1 | | | |
| 2002 | 5.08 | 123.11 | 648.22 | 3205.57 | 2967.35 | 76145.52 | 2606127 |
| 2005 | 5.05 | 154.78 | 726.93 | 3567.42 | 3315.17 | 87647.08 | 2900315 |
| 2008 | 6.36 | 192.71 | 815.27 | 3964.97 | 3703.50 | 97635.38 | 3223519 |
| 2011 | 5.28 | 200.54 | 888.78 | 4391.11 | 3956.09 | 105640.27 | 3569971 |
| 2014 | 5.92 | 234.55 | 938.87 | 4836.36 | 4375.39 | 119489.61 | 3931964 |
| 2017 | 6.47 | 264.22 | 1011.01 | 5280.82 | 4817.21 | 130562.56 | 4293308 |
| Year | | | | Group 2 | | | |
| 2003 | 4.72 | 129.87 | 675.30 | 3322.05 | 3071.13 | 80892.95 | 2700823 |
| 2006 | 5.55 | 152.12 | 772.46 | 3697.05 | 3471.87 | 90774.191 | 3005699 |
| 2009 | 6.22 | 192.17 | 818.35 | 4105.32 | 3822.74 | 101780.14 | 3337627 |
| 2012 | 5.38 | 204.07 | 906.38 | 4538.33 | 4001.58 | 111069.02 | 3689662 |
| 2015 | 6.66 | 236.51 | 957.69 | 4985.26 | 4490.60 | 123155.64 | 4053015 |
| 2018 | 6.92 | 274.25 | 1056.41 | 5428.77 | 4907.74 | 133395.06 | 4413589 |

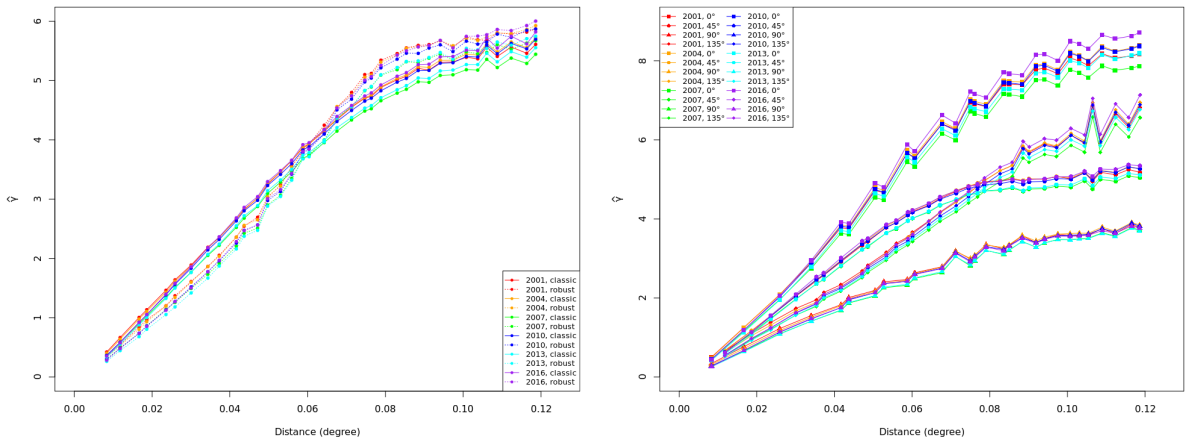
counts within the territory of county of Nairobi. More considerable increases in population counts could be observed in several very populous areas, and steady increases in the areal sizes of these very populous areas could be found as well.

A plot of the classic and robust versions of empirical semivariogram and directional empirical semivariogram values for the WorldPop population data at a 30 arc-seconds resolution for the years of 2001, 2004, 2007, 2010, 2013 and 2016 (i.e., the estimation group) on the transformed scale $g_3(y_i)$ is shown in Figure 6.2. Generally, the assumption of second-order stationarity, mentioned in the basic spatial modelling section of Chapter 3, is found acceptable, because the semivariograms of those population data were almost bounded (Atkinson and Lloyd, 2009), while the assumption of isotropy, as mentioned in the basic spatial modelling section of Chapter 3, has to be spoiled due to the bifurcation of the semivariograms in four different directions (Bivand et al., 2008). It is expected that the small amount of data extracted from Nairobi could reduce computation costs, even if the small areal size of input data and the unbalanced minimum N-S and W-E distances leads to anisotropy.

Box plots for the estimation group and two prediction groups are provided as Figure 6.3 to show the temporary variations of population counts on the transformed scale $g_3(y_i)$ within the territory of county of Nairobi. A steady overall increasing trend could be therefore observed.

## 6.4 Results

Figure 6.2a suggests the range of the semivariograms were around 0.09 degree for all those population data in the estimation group, so a sub-0.09-degree largest allowed triangle edge length is natural to consider. Consequently, the final chosen largest allowed triangle edge lengths are 0.02 degree for the inner area, and 0.04 degree for the outer area. A cut-off on edge length of 0.01 degree is again used. The theoretical basis of applying such specifications




(a) (b)

Figure 6.2: Classic and robust versions of (a) the empirical variograms and (b) the directional empirical variograms of the WorldPop data for the year of 2001, 2004, 2007, 2010, 2013 and 2016 at a resolution of 30 arc-seconds in red, orange, green, blue, cyan and purple on the transformed scale $g_3(y_i)$ (i.e., the estimation group).

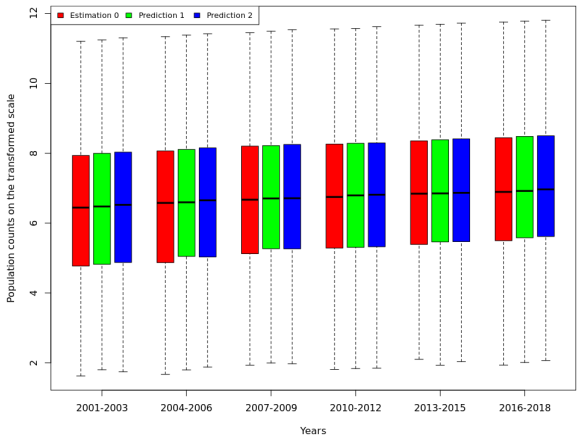



Figure 6.3: Box plots of population counts on the transformed scale $g_3(y_i)$ for the estimation group in red, the prediction group 1 (i.e., Group 1) in green and the prediction group 2 (i.e., Group 2) in blue.

is already discussed in the results section of Chapter 4. The triangular mesh employed for the calculation, mentioned in Chapter 6, can then be visualised as Figure 6.4.

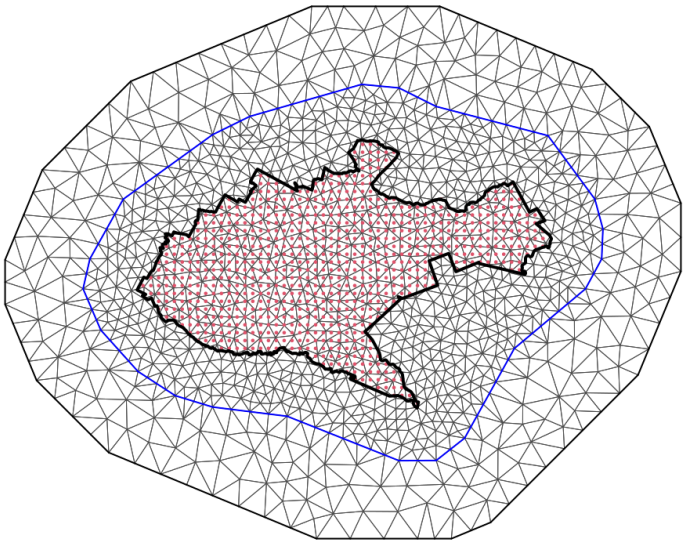


Figure 6.4: Triangular mesh used for spatiotemporal modelling with estimation locations (i.e., the locations where the WorldPop data at a 30 arc-seconds resolution in the estimation group lay) indicated by red dots.

Following the procedures as introduced in the research design section of this chapter, with the experience of transformation selection slightly spoiled, the estimation was made with all the WorldPop data in the estimation group, transformed with $g_3(y_i)$, at a 30 arc-seconds resolution and associated NTL and slope data. Vague priors were set for each hyperparameters (i.e., $\theta_{01}$ and $\theta_{02}$ associated with $\xi_0$, $\theta_{11}$ and $\theta_{12}$ associated with $\xi_1$, $\ln((1+\rho_0)/(1-\rho_0))$, $\ln((1+\rho_1)/(1-\rho_1))$ and $\ln(\sigma_e^{-2})$) on the internal scale used by `R-INLA` package and parameters (i.e., $\beta_0$, $\beta_1$ and $\beta_2$) for this exploration, as introduced in the priors section of Chapter 3 and summarised in Table 3.1. The posterior distributions of the hyperparameters (i.e., $\sigma_e^2$, $\sigma_0^2$ and $r_0$ associated with $\xi_0$, $\sigma_1^2$ and $r_1$ associated with $\xi_1$, $\rho_0$ and $\rho_1$) and parameters derived with the estimation are shown as Figure 6.5. It is observed that the temporal effects are small but present, and they are important for capturing spatial features of population growth in Nairobi (e.g., areal enlargement of satellite towns). Significance of the slope term is much higher than that of the NTL term in Nairobi, while it is contrary in Utrecht. This is possibly because electric power consumption per capita in Kenya (164 kWh) was much lower than that in the Netherlands (6713 kWh) as revealed by OECD/IEA (2014) and thus the NTL data less effectively reflect population counts and distributions. Besides, Ren et al. (2020) argued that the informal settlements around the 4 km belt in Nairobi has become the choice for most poor people, and that means the incomplete suburban infrastructure (e.g., NTL) may be not a strong indicator in such low-income settings. The effect of slope on indicating transformed population counts is positive, since millions of people live on the lower slopes of mountains in Kenya, close to main "water tower" forests (see Bussmann, 2002).

Figure C.3, Figure C.4, Figure C.5, and Figure C.6 available in Appendix C show the mapped posterior means and SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution

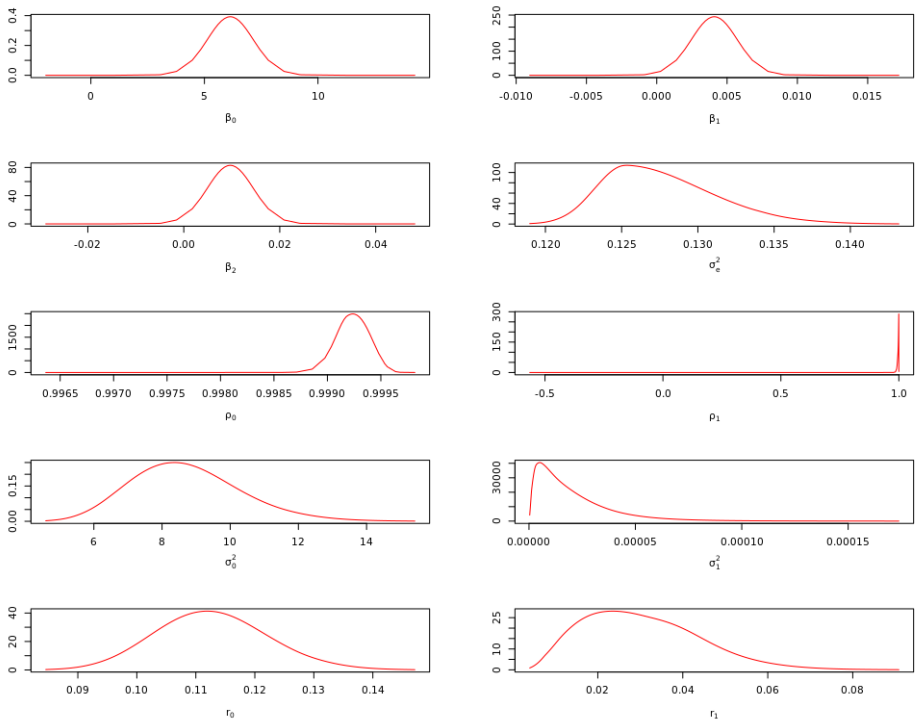



Figure 6.5: Posterior distributions of parameters and hyperparameters for estimations made with the WorldPop data in the estimation group transformed with $g_3(y_i)$ at a 30 arc-seconds resolution.

of 30 arc-seconds on the basis of the estimation made with all the WorldPop population data in the estimation group for prediction on Group 1 and Group 2. Table 6.2 showed the predicted population counts on the original scale per 30 arc-seconds gridded cell and their totals derived on the basis of the spatiotemporal estimation, comparable to the data in Table 6.1. Table 6.3 summarises the predictive performances of the spatiotemporal model at each time knot of each group for prediction with selection criteria PCC and RMSE derived by comparing predicted population counts on the original scale with retained WorldPop data. It is found that (i) areal enlargements of populous areas are well captured, but increments of population counts in populous areas are severely underestimated especially at the later time knots; (ii) obvious difference in qualities of predictions on two validation groups (i.e., Group 1 and Group 2) is not observed. As previously mentioned in the research design section of this chapter, the severe underestimations are probably caused by the use of the NTL data which are not a strong indicator for detecting the informally settled population in low-income settings. As mentioned in the research design section of this chapter, the effectiveness of such a temporal projection would largely depend on how similar the known population counts and distributions at adjacent time knots are. In this case, the predictions were made at the time knots that are not far away from the time knots where the estimation was actually made, and that means the known population data are already good approximations to the unknown ones at neighbouring time knots. Although, in this chapter, the temporal projection is only implemented in the discrete time domain, user of this approach may try to use it also in a continuous time domain in a real practice.

Table 6.2: Predicted population counts on the original scale per 30 arc-seconds gridded cell and their totals derived on the basis of the spatiotemporal estimation.

| WorldPop | Min | 25% | 50% | Mean | 75% | Max | Total |
|---|---|---|---|---|---|---|---|
| Year | | | | Group 1 | | | |
| 2002 | 4.78 | 132.21 | 554.09 | 2737.44 | 2637.15 | 41835.73 | 2225538 |
| 2005 | 5.14 | 143.96 | 614.33 | 2965.39 | 2910.89 | 44256.82 | 2410864 |
| 2008 | 6.75 | 171.52 | 717.71 | 3431.10 | 3397.52 | 53897.63 | 2789486 |
| 2011 | 6.34 | 180.82 | 798.39 | 3849.50 | 3794.53 | 64455.49 | 3129646 |
| 2014 | 7.9 | 209.2 | 870.5 | 4219.6 | 4182.3 | 68525.2 | 3430556 |
| 2017 | 7.25 | 225.86 | 939.23 | 4544.12 | 4550.50 | 74288.36 | 3694367 |
| Year | | | | Group 2 | | | |
| 2003 | 4.76 | 131.48 | 553.98 | 2700.90 | 2637.15 | 40993.25 | 2195836 |
| 2006 | 5.1 | 145.7 | 626.0 | 3018.6 | 2953.8 | 47677.0 | 2454079 |
| 2009 | 6.69 | 172.81 | 725.37 | 3506.77 | 3437.88 | 56899.34 | 2851006 |
| 2012 | 6.79 | 187.30 | 794.17 | 3797.01 | 3680.35 | 61040.84 | 3086972 |
| 2015 | 8.0 | 212.1 | 881.1 | 4253.8 | 4230.6 | 68993.6 | 3458373 |
| 2018 | 7.25 | 231.17 | 939.07 | 4553.65 | 4570.82 | 74797.23 | 3702114 |

Table 6.3: Predictive performances of the spatiotemporal model at each time knot of each group for prediction on the original scale $y_i$ in 30 arc-seconds grid cells.

| Criteria | | | | Group 1 | | |
|---|---|---|---|---|---|---|
| Year | 2002 | 2005 | 2008 | 2011 | 2014 | 2017 |
| PCC | 0.8723578 | 0.8642024 | 0.8701749 | 0.8677075 | 0.8630609 | 0.8644906 |
| RMSE | 3232.224 | 3764.037 | 4045.474 | 4516.232 | 5138.6 | 5556.85 |
| Criteria | | | | Group 2 | | |
| Year | 2003 | 2006 | 2009 | 2012 | 2015 | 2018 |
| PCC | 0.8666798 | 0.868584 | 0.8710506 | 0.8656544 | 0.8687612 | 0.8685694 |
| RMSE | 3487.669 | 3877.648 | 4197.232 | 4803.527 | 5218.512 | 5664.666 |

# Chapter 7

# Conclusion

This chapter aims to summarise the contents of the previous chapters, conclude the findings, suggest further improvements that could be done and summarise the contributions of this master's thesis to science.

## 7.1 Chapter-wise Summary

Chapter 1 introduces the importance of mapping population precisely at local levels and stresses the problem of spatial and temporal mismatches between the target applications and the spatial and temporal resolutions of the large-scale population data products can provide. It considers each large-scale gridded population data products based on "top-down" methods as unique and valuable sources of population data that are irreplaceable due to scarcity of population census and ancillary data, lack of standard dasymetric methods and different population definitions, and irreproducible due to proprietary. It also introduces the rising "bottom-up" methods based on the "microcensus" household surveys and previous efforts in refining population mapping, and discusses the strengths and drawbacks. It finally introduces four novel ideas proposed in this master's thesis: the idea of simplifying the use of ancillary data, the idea of projecting existing gridded population estimates to any combination of spatial and temporal resolutions, the idea of assessing the border between "top-down" and "bottom-up" approaches, and the idea of combining "top-down" and "bottom-up" approaches.

Chapter 2 introduces the background information of the GHS-POP data at a resolution of 30 arc-seconds (i.e., the ones to be spatially projected), GHS-POP data at a resolution of 15 arc-seconds (the ones used as the real population data in Chapter 4), CBS population data at a resolution of 15 arc-seconds (the ones used as the real population data in Chapter 5) and WorldPop data at a resolution of 30 arc-seconds (i.e., the ones to be temporally projected) at first. It also introduces the idea of simplifying the use of ancillary data, which refers to using the easy-to-update or hard-to-change ones but not the easy-to-change but hard-to-update ones, and that enables deriving prior information for the "bottom-up" models (i.e., the idea of combining "top-down" and "bottom-up" approaches).

Chapter 3 introduces the Bayesian spatial model and its implementation through the SPDE-INLA approach. The model's Bayesian nature allows the prior information derived with large-scale gridded population data products available for free and accessible the easy-to-update or hard-to-change ancillary data also available for free to be incorporated into the "bottom-up" models, so as to implement the idea of combining "top-down" and "bottom-up" approaches.

The capacity of the Matérn-parameters-based spatial model in capturing the spatial autocorrelation inherent in data is stronger than that of the hierarchical random intercept proposed by Leasure et al. (2020), while the use of the INLA approach reduces the burden of computation costs.

In Chapter 4, the idea of assessing the border between "top-down" and "bottom-up" approaches is tested with the GHS-POP data at a resolution of 15 arc-seconds treated as the real population data. The basis is the idea of projecting existing gridded population estimates, referring to testing whether the large-scale gridded population estimates based on "top-down" methods could be projected with the NTL and DEM data to a target spatial resolution that existing data products have not provided. The idea of assessing the border between "top-down" and "bottom-up" approaches refers to testing whether the projected population estimates have better predictive performances than the ones acquired in a "bottom-up" approach. The "bottom-up" models are established with samples of different sizes, randomly drawn with or without weights from the real population data, as the likelihood and existing gridded population estimates as the prior information, delivered through different types of priors that reflect different levels of informativeness.

In Chapter 5, the idea of assessing the border between "top-down" and "bottom-up" approaches is tested with the CBS population data at a resolution of 15 arc-seconds treated as the real population data. This time the real population data directly reflect true information about population distributions and counts, while the real population data defined in Chapter 4 are just estimates on the real population data. This chapter indeed considers the measurement errors and statistical errors contained in the existing gridded population estimates based on "top-down" methods.

Chapter 6 extends the Bayesian spatial model to the Bayesian spatiotemporal model. The idea of projecting existing gridded population estimates now refers to testing whether the large-scale gridded population estimates based on "top-down" methods could be projected with the NTL and DEM data to a target temporal resolution that existing data products have not provided.

R code for the proposed approach is available in a GitHub repository (`https://github.com/yueyangyi/kul_thesis_model_final/blob/main/kul_thesis_model_final.R`).

## 7.2 Findings and Recommendations

The core ideas generally function well. A few conclusive suggestions could be made to the user of the proposed approach: (i) a dispersion parameter (e.g., unstructured variance $\sigma_e^2$) is important for modelling population counts; a model without a dispersion parameter (e.g., the LGCP model) suffers a serious underdispersion or overdispersion problem; (ii) the Matérn-parameters-based geostatistical model can more effectively describe the spatial characteristics of population distributions and counts than the hierarchical-random-intercept-based non-spatial model proposed by Leasure et al. (2020), thanks to the development of the INLA approach; (iii) the transformation function $g_3(y_i)$ is a safe option for avoiding some very large posterior means or SD of unstructured variance $\sigma_e^2$ and structured variance $\sigma^2$; (iv) the estimations based weighted sampling usually have better predictive performances than the ones based on the unweighted sampling; (v) given the large-scale gridded population data products based on "top-down" methods and available for free provide very accurate estimates on the real popu-

lation distributions and counts, the informative normal priors should be used to improve the underestimation and overestimation of population counts; (vi) given the large-scale gridded population data products cannot provide very accurate estimates on the real population distributions and counts, the weakly informative PC priors should be used for their mostly harmless nature; (vii) a meaningful "bottom-up" "microcensus" household survey does not necessarily means achieving good predictive performances, and how nice a "free lunch" is and how well a "bottom-up" model can perform largely depend on how well the large-scale gridded population data products estimate the real population distributions and counts; (viii) the effectiveness of a temporal projection would largely depend on how similar the known population counts and distributions at adjacent time knots are.

A few suggestions on further improvements could be made as follows: (i) in Chapter 4 and Chapter 6, it is mentioned that the harmonised DMSP-like NTL data suffer from the "overglow" effect so cannot be regarded as a good indicator of population presence in less populous areas; however, the successor of DMSP-OLS data, VIIRS-DNB data, only have a short time series from 2012 to 2020; the data indeed minimise the "overglow" effect and and bright saturation that compromise DMSP-OLS composites (Li et al., 2020) but cannot meanwhile describe the saturation of population counts in the most populous areas (i.e., the use of VIIRS-DNB data would result in extremely large predicted population counts at the original scale in the most populous areas); Hence, a preprocessing approach that defuses the reduction on bright saturation in the most populous areas could be developed in the future; (ii) in a real practice, when the real population data (e.g., the CBS population data used in Chapter 5) are not available, a test with existing "top-down" population estimates (e.g., the GHS-POP data used in Chapter 4) defined as the real population data would be the sole one that could suggest a suitable sample size for the "bottom-up" approach, and this idea could be combined with the approach proposed by Boo et al. (2020), which refers to finding out the acceptable range of sample size of the "microcensus" with the Kolmogorov-Smirnov distance between (weighted and/or unweighted) empirical cumulative distribution function of population counts acquired from the large-scale population data products with different sample sizes and empirical cumulative distribution function of all population counts that can be sampled; the an acceptable region of sample size, restricted within these boundaries, could then be found; (iii) the simulation study conducted by Leasure et al. (2021) demonstrates that population-weighted random sampling results in biased estimates of population densities, but it is still unclear whether the spatial model proposed in this master's thesis is sensitive to this bias; an approach that recovers unbiased population estimates from weighted survey data may be required; (iv) the transformation function $g_3(y_i)$ functions well in this master's thesis, but it would also bring a bias likely to be endemic, as demonstrated by Cohn et al. (2022); whether the spatial model proposed in this master's thesis is sensitive to this bias could also be investigated in the future; (v) the proposed approach could be regarded as a general solution to the dispersion problem, while upgrading the Poisson distribution used in the LGCP model to a generalised Poisson distribution could be another; the generalised Poisson distribution defined by Consul and Jain (1973) has flexibility in dealing with the overdispersion problem relative to the Poisson distribution, but it is unable to capture some levels of underdispersion (Shmueli et al., 2005); the Conway-Maxwell-Poisson distribution proposed by Conway and Maxwell (1961) also generalises the Poisson distribution, and it is found to be flexible for fitting overdispersed and underdispersed data (Shmueli et al., 2005) and even able to model arbitrarily underdispersed counts (Huang, 2023); the Conway-Maxwell-Poisson distribution has not yet been included in

the R-INLA package, but it is certainly worth a test in the future.

## 7.3 Contributions to Science

Accurate regional population mapping at local levels is important for researchers to conduct novel population-related research and for policymakers to practice population-related policies. This master's thesis develops an approach that projects existing gridded population estimates based on "top-down" methods and available for free to any target pair of spatial and temporal resolutions to facilitate a broader range of applications, with assistance from easy-to-measure or hard-to-change geospatial ancillary data. If the projected "top-down" population estimates based on the existing gridded population data products available for free hold the same or even better quality than the ones achieved in a "bottom-up" approach based on the limited small-scale "microcensus" household surveys, a "free lunch" could be enjoyed. It means, in such a situation, the user does not have to pay for the "microcensus", as paying extra does not bring about better predictive performances. This idea is the first attempt trying to assess the border between "top-down" and "bottom-up" methods, so as to find the minimum sample size that lets the predictive performances of "bottom-up" estimates be better than the projected large-scale population data products available for free. It is also the first time to use ancillary data available for free and "top-down" global population data products available also for free for calculating prior information for "bottom-up" estimates. The proposed approach represents a significant step towards a combination of mainstream "top-down" and rising "bottom-up" approaches that makes the most of census and survey data.

# Bibliography


Adde, A., Darveau, M., Barker, N., Cumming, S., and López, A. B. (2020). Predicting spatiotemporal abundance of breeding waterfowl across canada: A bayesian hierarchical modelling approach. *Diversity and Distributions*, 26(10):1248–1263.

Archila Bustos, M. F., Hall, O., Niedomysl, T., and Ernstson, U. (2020). A pixel level evaluation of five multitemporal global gridded population datasets: a case study in sweden, 1990-2015. *Population and Environment*, 42(2):255–277.

Atkinson, P. M. and Lloyd, C. D. (2009). Geostatistics and spatial interpolation. In *The SAGE Handbook of Spatial Analysis*, chapter 9, pages 159–181. SAGE Publications Ltd, London.

Bagan, H. and Yamagata, Y. (2015). Analysis of urban growth and estimating population density using satellite images of nighttime lights and land-use and population data. *GIScience & Remote Sensing*, 52(6):765–780.

Balk, D., Pozzi, F., Yetman, G., Deichmann, U., and Nelson, A. (2005). *The distribution of people and the dimension of place: methodologies to improve the global estimation of urban extents*. Proceedings of the Urban Remote Sensing Conference. International Society for Photogrammetry and Remote Sensing, Tempe, AZ.

Batista e Silva, F., Dijkstra, L., and Poelman, H. (2021). *The JRC-GEOSTAT 2018 population grid*. Joint Research Centre (JRC), European Commission.

Bivand, R. S. (2010). Exploratory spatial data analysis. In Fischer, M. M. and Getis, A., editors, *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, pages 219–254. Springer Berlin Heidelberg, Berlin, Heidelberg.

Bivand, R. S., Pebesma, E. J., and Gómez-Rubio, V. (2008). Interpolation and geostatistics. In *Applied Spatial Data Analysis with R*, pages 191–235. Springer New York, New York, NY.

Blangiardo, M. and Cameletti, M. (2015). *Spatial and Spatio-temporal Bayesian Models with R-INLA*. John Wiley & Sons, Chichester.

Boo, G., Darin, E., Leasure, D. R., Dooley, C. A., Chamberlain, H. R., Lazar, A. N., Tschirhart, K., Sinai, C., Hoff, N. A., Fuller, T., Musene, K., Batumbo, A., Rimoin, A. W., and Tatem, A. J. (2022). High-resolution population estimation using household survey data and building footprints. *Nature Communications*, 13(1):1330.

Boo, G., Darin, E., Thomson, D. R., and Tatem, A. J. (2020). A grid-based sample design framework for household surveys. *Gates Open Research*, 4(13).

Bussmann, R. W. (2002). Islands in the desert-forest vegetation of kenya's smaller mountains and highland areas (nyiru, ndoto, kulal, marsabit, loroghi, ndare, mukogodo, porror, mathews, gakoe, imenti, ngaia, nyambeni, loita, nguruman, nairobi). *Journal of East African Natural History*, 91(1):27–79.

Carleton, T., Cornetet, J., Huybers, P., Meng, K. C., and Proctor, J. (2021). Global evidence for ultraviolet radiation decreasing covid-19 growth rates. *Proceedings of the National Academy of Sciences*, 118(1):e2012370118.

CBS (2015). *CBS gebiedsindelingen*. Centraal Bureau voor de Statistiek.

CBS (2019). *Kaart van 100 meter bij 100 meter met statistieken*. Centraal Bureau voor de Statistiek.

CBS (2023). *StatLine: Bevolkingsontwikkeling; regio per maand*. Centraal Bureau voor de Statistiek.

Cohn, J. B., Liu, Z., and Wardlaw, M. I. (2022). Count (and count-like) data in finance. *Journal of Financial Economics*, 146(2):529–551.

Consul, P. C. and Jain, G. C. (1973). A generalization of the poisson distribution. *Technometrics*, 15(4):791–799.

Conway, R. W. and Maxwell, W. L. (1961). A queuing model with state dependent service rates. *The Journal of Industrial Engineering*, 12(2):132–136.

Cressie, N. A. C. (1993). Geostatistics. In *Statistics for Spatial Data*, chapter 2, pages 27–104. John Wiley & Sons, Ltd.

Czado, C., Gneiting, T., and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, 65(4):1254–1261.

Dawid, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, 147(2):278–292.

Doxsey-Whitfield, E., MacManus, K., Adamo, S. B., Pistolesi, L., Squires, J., Borkovska, O., and Baptista, S. R. (2015). Taking advantage of the improved availability of census data: A first look at the gridded population of the world, version 4. *Papers in Applied Geography*, 1(3):226–234.

Elvidge, C., Hsu, F. C., Baugh, K., and Gosh, T. (2014). *National trends in satellite-observed lighting 1992-2012*. Global urban monitoring and assessment through earth observation. CRC Press, Boca Raton.

Forlani, C., Bhatt, S., Cameletti, M., Krainski, E., and Blangiardo, M. (2020). A joint bayesian space-time model to integrate spatially misaligned air pollution data in r-inla. *Environmetrics*, 31(8).

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2019). Constructing priors that penalize the complexity of gaussian random fields. *Journal of the American Statistical Association*, 114(525):445–452.

Gaughan, A. E., Stevens, F. R., Huang, Z., Nieves, J. J., Sorichetta, A., Lai, S., Ye, X., Linard, C., Hornby, G. M., Hay, S. I., Yu, H., and Tatem, A. J. (2016). Spatiotemporal patterns of population in mainland china, 1990 to 2010. *Scientific Data*, 3:160005.

Gelman, A. (2007). Struggles with Survey Weighting and Regression Modeling. *Statistical Science*, 22(2):153 – 164.

Giani, P., Castruccio, S., Anav, A., Howard, D., Hu, W., and Crippa, P. (2020). Short-term and long-term health impacts of air pollution reductions from covid-19 lockdowns in china and europe: a modelling study. *The Lancet Planetary Health*, 4(10):e474–e482.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Haug, O., Thorarinsdottir, T. L., Sørbye, S. H., and Franzke, C. L. E. (2020). Spatial trend analysis of gridded temperature data at varying spatial scales. *Advances in Statistical Climatology, Meteorology and Oceanography*, 6(1):1–12.

Henderson, J. V., Nigmatulina, D., and Kriticos, S. (2021). Measuring urban economic density. *Journal of Urban Economics*, 125:103188.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15):1965–1978.

Huang, A. (2023). On arbitrarily underdispersed discrete distributions. *The American Statistician*, 77(1):29–34.

Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., and Rue, H. (2019). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. CRC Press, Boca Raton.

Leasure, D. R., Dooley, C. A., and Tatem, A. J. (2021). *A simulation study exploring weighted Bayesian models to recover unbiased population estimates from weighted survey data*. WorldPop, University of Southampton.

Leasure, D. R., Jochem, W. C., Weber, E. M., Seaman, V., and Tatem, A. J. (2020). National population mapping from sparse survey data: A hierarchical bayesian modeling framework to account for uncertainty. *Proceedings of the National Academy of Sciences*, 117(39):24173–24179.

Leyk, S., Gaughan, A. E., Adamo, S. B., de Sherbinin, A., Balk, D., Freire, S., Rose, A., Stevens, F. R., Blankespoor, B., Frye, C., Comenetz, J., Sorichetta, A., MacManus, K., Pistolesi, L., Levy, M., Tatem, A. J., and 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.

Li, X., Zhou, Y., Zhao, M., and Zhao, X. (2020). A harmonized global nighttime light dataset 1992-2018. *Scientific Data*, 7(1):168.

Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with r-inla. *Journal of Statistical Software*, 63(19):1 – 25.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Lloyd, C. T., Chamberlain, H., Kerr, D., Yetman, G., Pistolesi, L., Stevens, F. R., Gaughan, A. E., Nieves, J. J., Hornby, G., MacManus, K., Sinha, P., Bondarenko, M., Sorichetta, A., and Tatem, A. J. (2019). Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets. *Big Earth Data*, 3(2):108–139.

Lu, D., Wang, Y., Yang, Q., Su, K., Zhang, H., and Li, Y. (2021). Modeling spatiotemporal population changes by integrating dmsp-ols and npp-viirs nighttime light data in chongqing, china. *Remote Sensing*, 13(2):284.

Maclaurin, G., Leyk, S., and Hunter, L. (2015). Understanding the combined impacts of aggregation and spatial non-stationarity: The case of migration-environment associations in rural south africa. *Transactions in GIS*, 19(6):877–895.

MacManus, K., Balk, D., Engin, H., McGranahan, G., and Inman, R. (2021). Estimating population and urban areas at risk of coastal hazards, 1990-2015: how data choices matter. *Earth System Science Data*, 13(12):5747–5801.

Mohanty, M. P. and Simonovic, S. P. (2021). Understanding dynamics of population flood exposure in canada with multiple high-resolution population datasets. *Science of The Total Environment*, 759:143559.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log gaussian cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482.

Nagle, N. N., Buttenfield, B. P., Leyk, S., and Spielman, S. (2014). Dasymetric modeling and uncertainty. *Annals of the Association of American Geographers*, 104(1):80–95.

Ndeng'e, G., Opiyo, C., Mistiaen, J., and Kristjanson, P. (2003). *Geographic dimensions of well-being in Kenya: Where are the poor? From districts to locations*. Central Bureau of Statistics, Kenya Ministry of Planning and National Development, Nairobi.

OECD/IEA (2014). *Electric power consumption (kWh per capita)*. International Energy Agency.

Pesaresi, M. and Politis, P. (2022). *GHS built-up surface grid, derived from Sentinel2 composite and Landsat, multitemporal (1975-2030)*. Joint Research Centre (JRC), European Commission.

QGIS Development Team (2022). *QGIS Geographic Information System*. QGIS Association.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ren, H., Guo, W., Zhang, Z., Kisovi, L. M., and Das, P. (2020). Population density and spatial patterns of informal settlements in nairobi, kenya. *Sustainability*, 12(18).

RStudio Team (2022). *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, Boca Raton.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Schiavina, M., Freire, S., and MacManus, K. (2019). *GHS population grid multitemporal (1975-1990-2000-2015), R2019A*. Joint Research Centre (JRC), European Commission.

Schiavina, M., Melchiorri, M., and Pesaresi, M. (2022). *GHS-SMOD R2022A - GHS settlement layers, application of the Degree of Urbanisation methodology (stage I) to GHS-POP R2022A and GHS-BUILT-S R2022A, multitemporal (1975-2030)*. Joint Research Centre (JRC), European Commission.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the conway-maxwell-poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):127–142.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. *Statistical Science*, 32(1):1 – 28.

Sinha, P., Gaughan, A. E., Stevens, F. R., Nieves, J. J., Sorichetta, A., and Tatem, A. J. (2019). Assessing the spatial sensitivity of a random forest model: Application in gridded population modeling. *Computers, Environment and Urban Systems*, 75:132–145.

Stathakis, D. and Baltas, P. (2018). Seasonal population estimates based on night-time lights. *Computers, Environment and Urban Systems*, 68:133–141.

Stevens, F. R., Gaughan, A. E., Linard, C., and Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLOS ONE*, 10(2):e0107042.

Sun, W., Zhang, X., Wang, N., and Cen, Y. (2017). Estimating population density using dmsp-ols night-time imagery and land cover data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(6):2674–2684.

Thomson, D. R., Rhoda, D. A., Tatem, A. J., and Castro, M. C. (2020). Gridded population survey sampling: a systematic scoping review of the field and strategic research agenda. *International Journal of Health Geographics*, 19(1):34.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., and Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1):1.

Wang, L., Wang, S., Zhou, Y., Liu, W., Hou, Y., Zhu, J., and Wang, F. (2018). Mapping population density in china between 1990 and 2010 using remote sensing. *Remote Sensing of Environment*, 210:269–281.

Ward, P. J., Blauhut, V., Bloemendaal, N., Daniell, J. E., de Ruiter, M. C., Duncan, M. J., Emberson, R., Jenkins, S. F., Kirschbaum, D., Kunz, M., Mohr, S., Muis, S., Riddell, G. A., Schäfer, A., Stanley, T., Veldkamp, T. I. E., and Winsemius, H. C. (2020). Review article: Natural hazard risk assessments at the global scale. *Natural Hazards and Earth System Sciences*, 20(4):1069–1096.

Ward, P. J., Jongman, B., Weiland, F. S., Bouwman, A., van Beek, R., Bierkens, M. F. P., Ligtvoet, W., and Winsemius, H. C. (2013). Assessing flood risk at the global scale: model setup, results, and sensitivity. *Environmental Research Letters*, 8(4):044019.

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

Warnes, J. J. and Ripley, B. D. (1987). Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika*, 74(3):640–642.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 41(3/4):434–449.

Wilson, K. and Wakefield, J. (2021). Estimation of health and demographic indicators with incomplete geographic information. *Spatial and Spatio-temporal Epidemiology*, 37:100421.

Winsemius, H. C., Van Beek, L. P. H., Jongman, B., Ward, P. J., and Bouwman, A. (2013). A framework for global river flood risk assessments. *Hydrology and Earth System Sciences*, 17(5):1871–1892.

WorldPop (2022). *Global High Resolution Population Denominators Project*. University of Southampton.

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., Sampson, C. C., Kanae, S., and Bates, P. D. (2017). A high-accuracy map of global terrain elevations. *Geophysical Research Letters*, 44(11):5844–5853.

Yu, S., Zhang, Z., and Liu, F. (2018). Monitoring population evolution in china using time-series dmsp/ols nightlight imagery. *Remote Sensing*, 10(2):194.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.

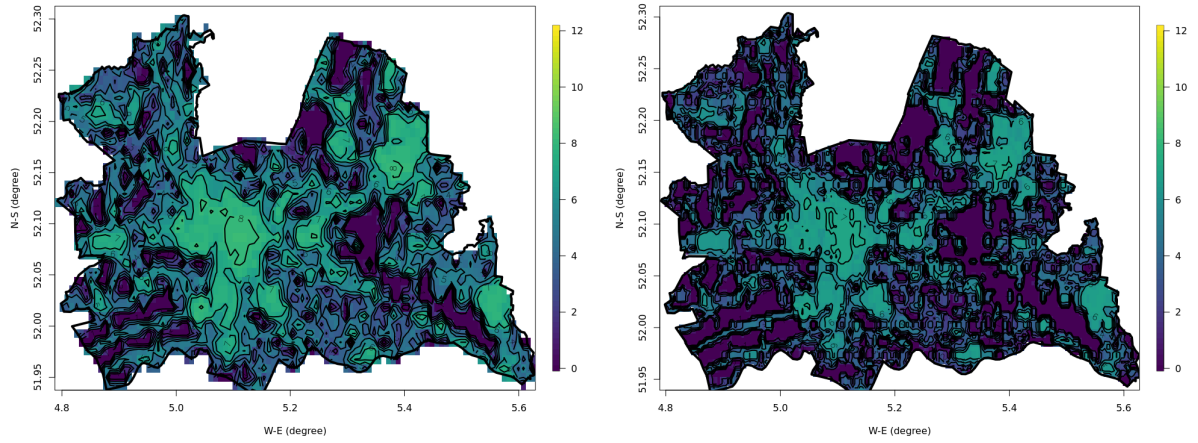
Zhao, N., Liu, Y., Cao, G., Samson, E. L., and Zhang, J. (2017). Forecasting china's gdp at the pixel level using nighttime lights time series and population images. *GIScience & Remote Sensing*, 54(3):407–425.

# Appendices

# Appendix A

# Appendix for Chapter 4

The GHS-POP data at resolutions of 30 arc-seconds (i.e., the population data to be projected) and 15 arc-seconds (i.e., the real population data in Chapter 4) are visualised as Figure A.1, with the transformation function $g_3(y_i)$ applied in order to show the patterns more clearly.




(a) (b)

Figure A.1: The GHS-POP data at resolutions of (a) 30 arc-seconds (i.e., the population data to be projected) and (b) 15 arc-seconds (i.e., the real population data defined in Chapter 4) on the transformed scale $g_3(y_i)$.

Figure A.2 and Figure A.3, Figure A.4 and Figure A.5, Figure A.6 and Figure A.7, Figure A.8 and Figure A.9, Figure A.10 and Figure A.5, and Figure A.12 and Figure A.13 show the mapped posterior means and SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of the thirty-six estimations made with (i) unweighted sampling and normal priors, (ii) unweighted sampling and PC priors, (iii) unweighted sampling and vague priors, (iv) weighted sampling and normal priors, (v) weighted sampling and PC priors, and (vi) weighted sampling and vague priors, and different sample sizes.

Figure A.14 and Figure A.15 present the mapped posterior means and SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of the estimations made with "bottom-up" sample of 100% full real population data and the ones derived by projecting existing "top-down" population estimates.

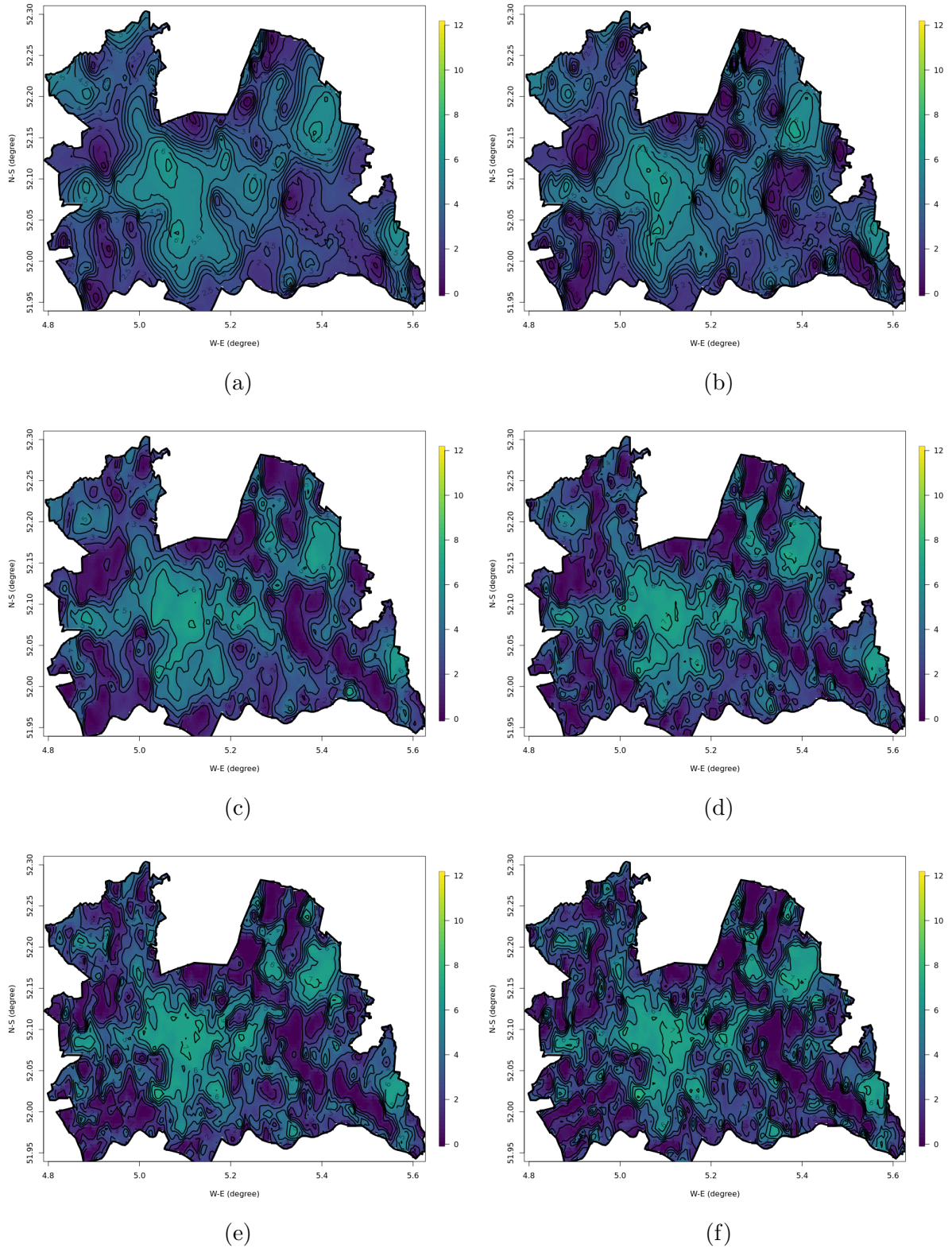



Figure A.2: Posterior mean of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with existing "top-down" estimates as normal priors and unweighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

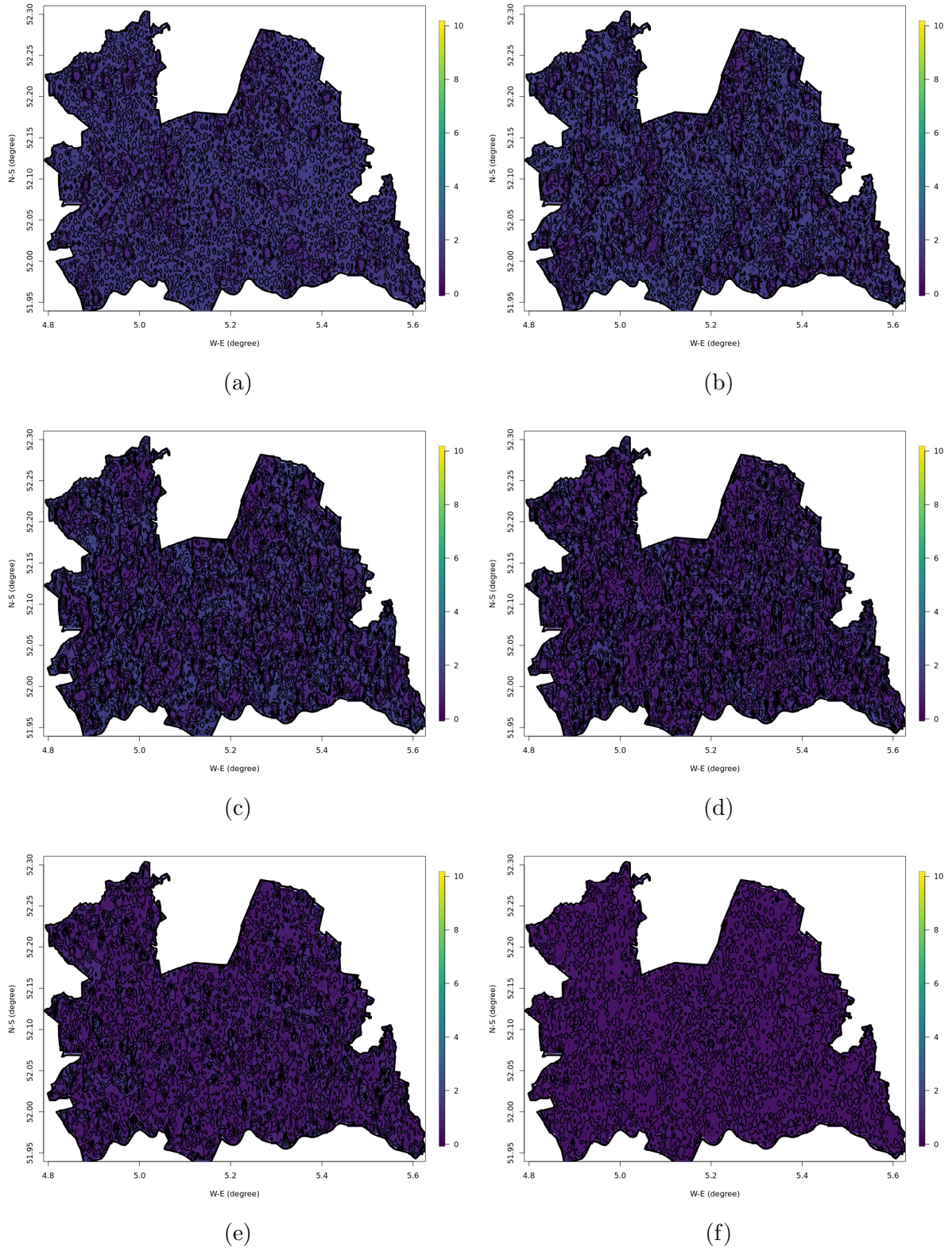



Figure A.3: Posterior SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with existing "top-down" estimates as normal priors and unweighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

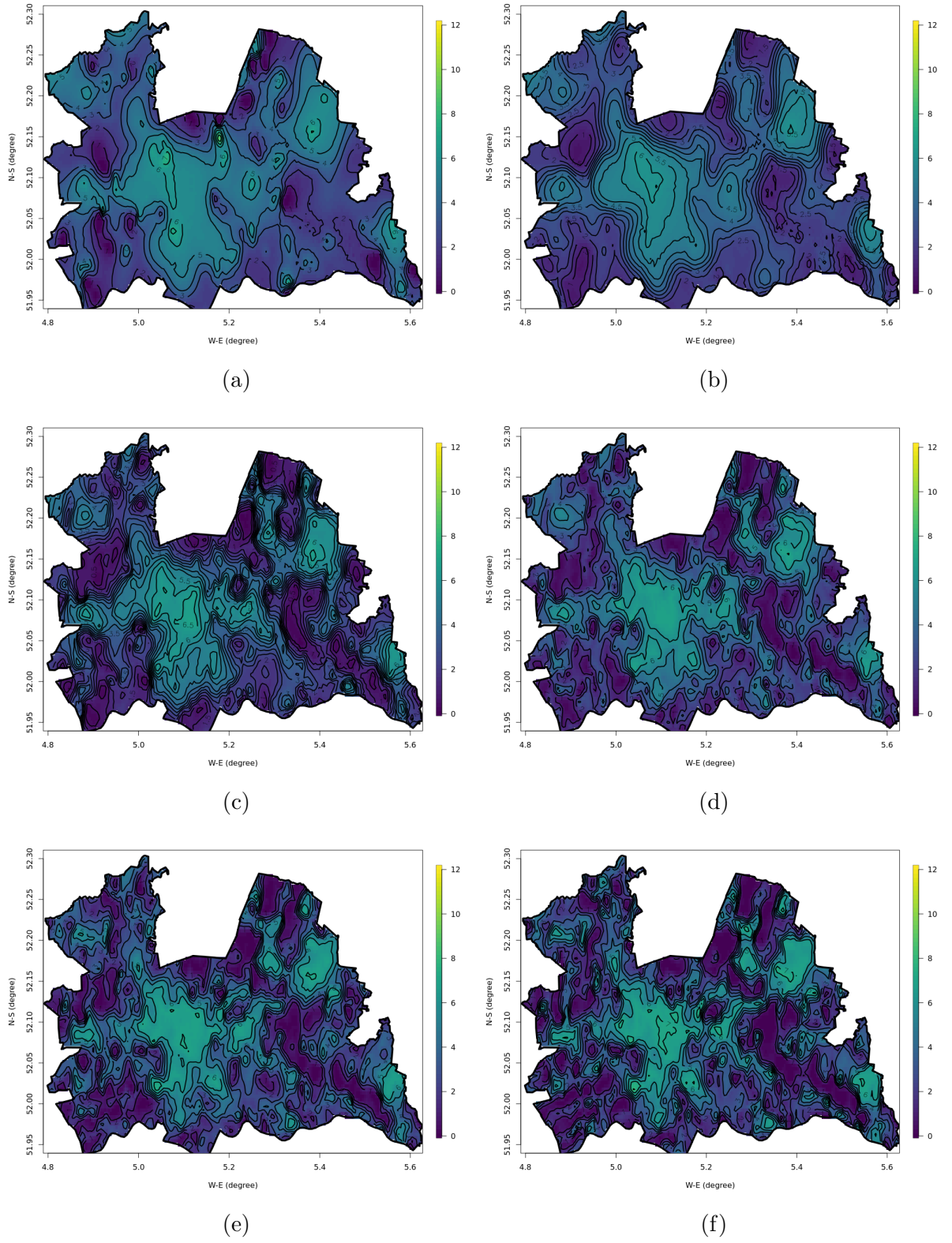



Figure A.4: Posterior mean of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with existing "top-down" estimates as PC priors and unweighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

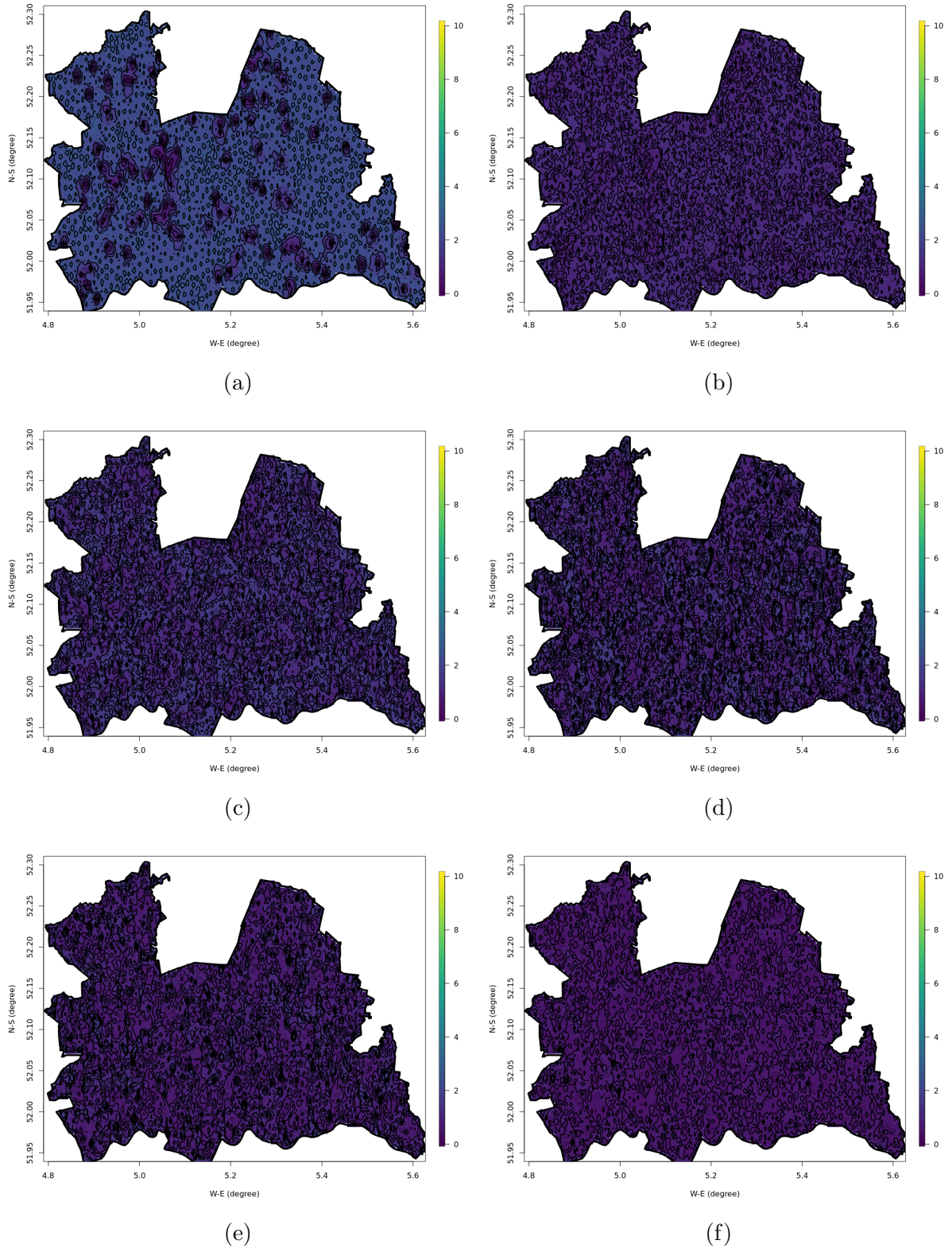



Figure A.5: Posterior SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with existing "top-down" estimates as PC priors and unweighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

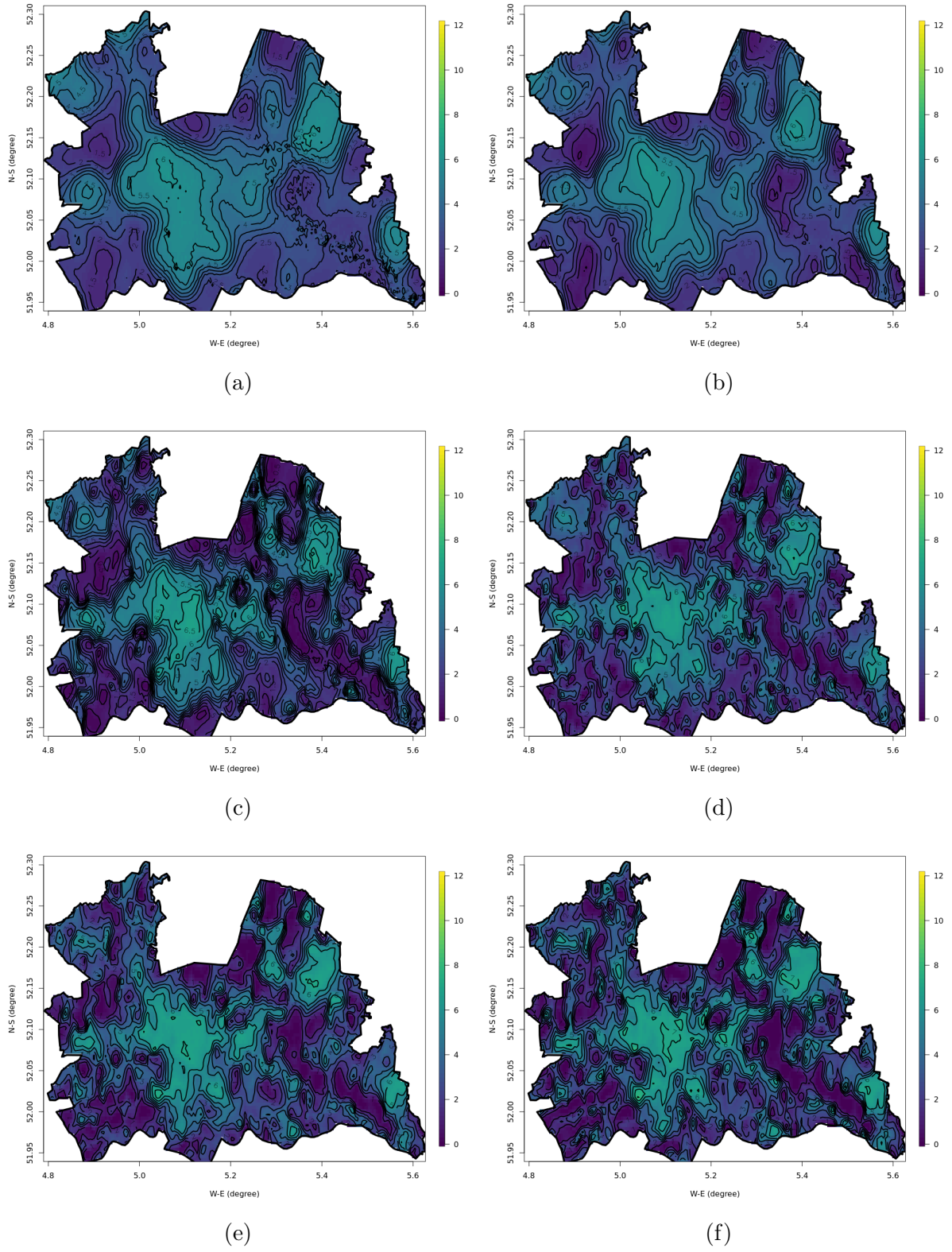



Figure A.6: Posterior mean of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with vague priors and unweighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

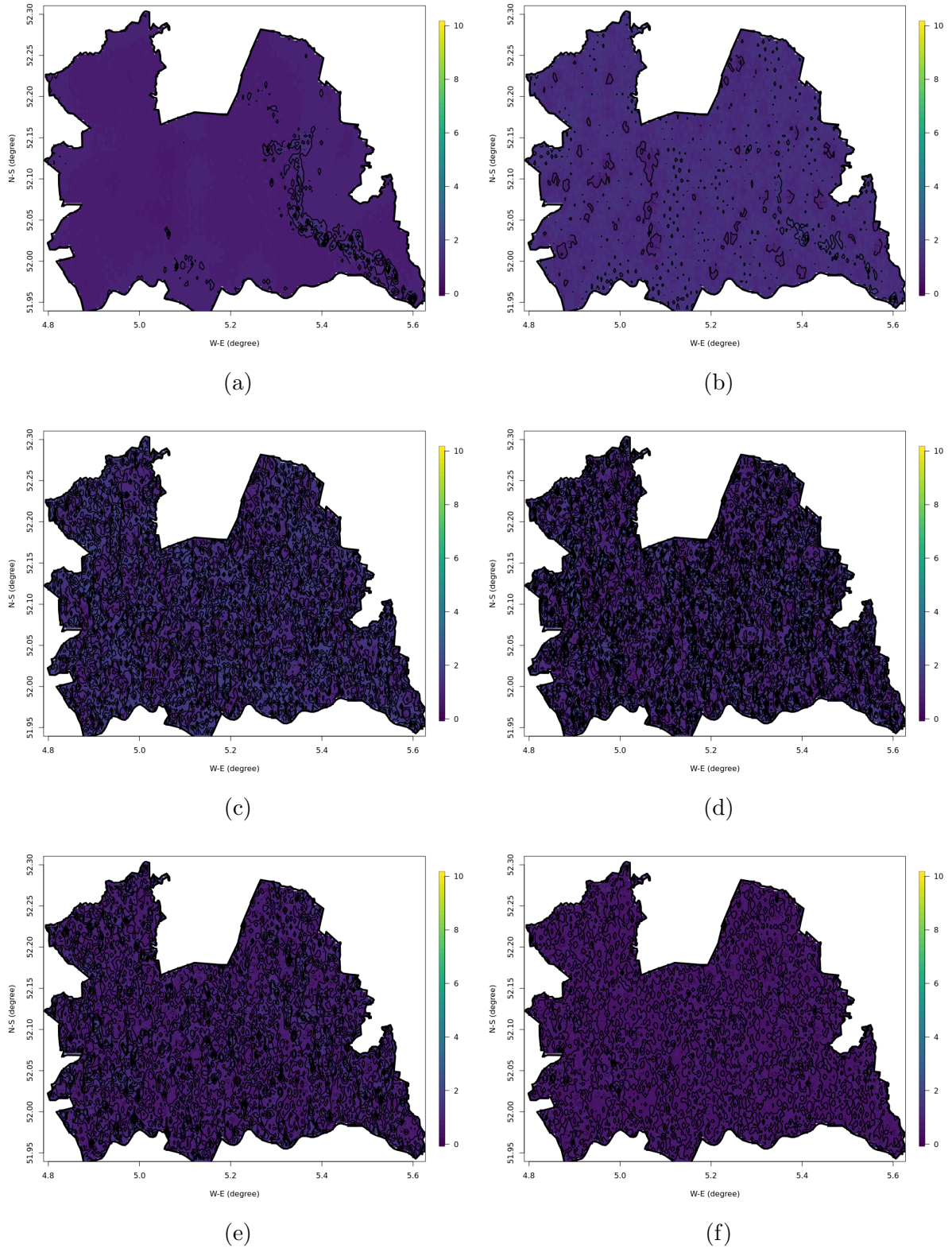



Figure A.7: Posterior SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with vague priors and unweighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

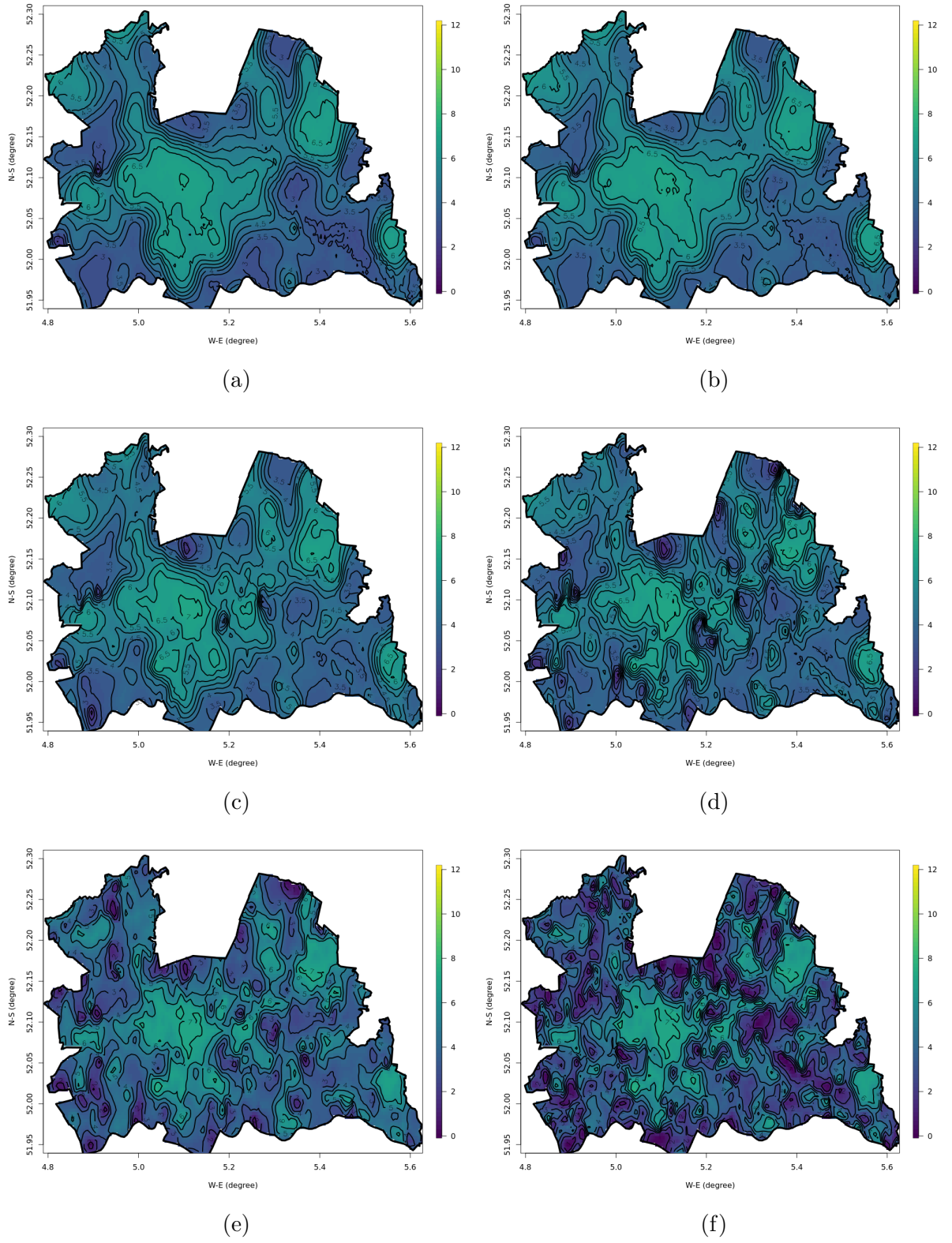



Figure A.8: Posterior mean of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with existing "top-down" estimates as normal priors and weighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

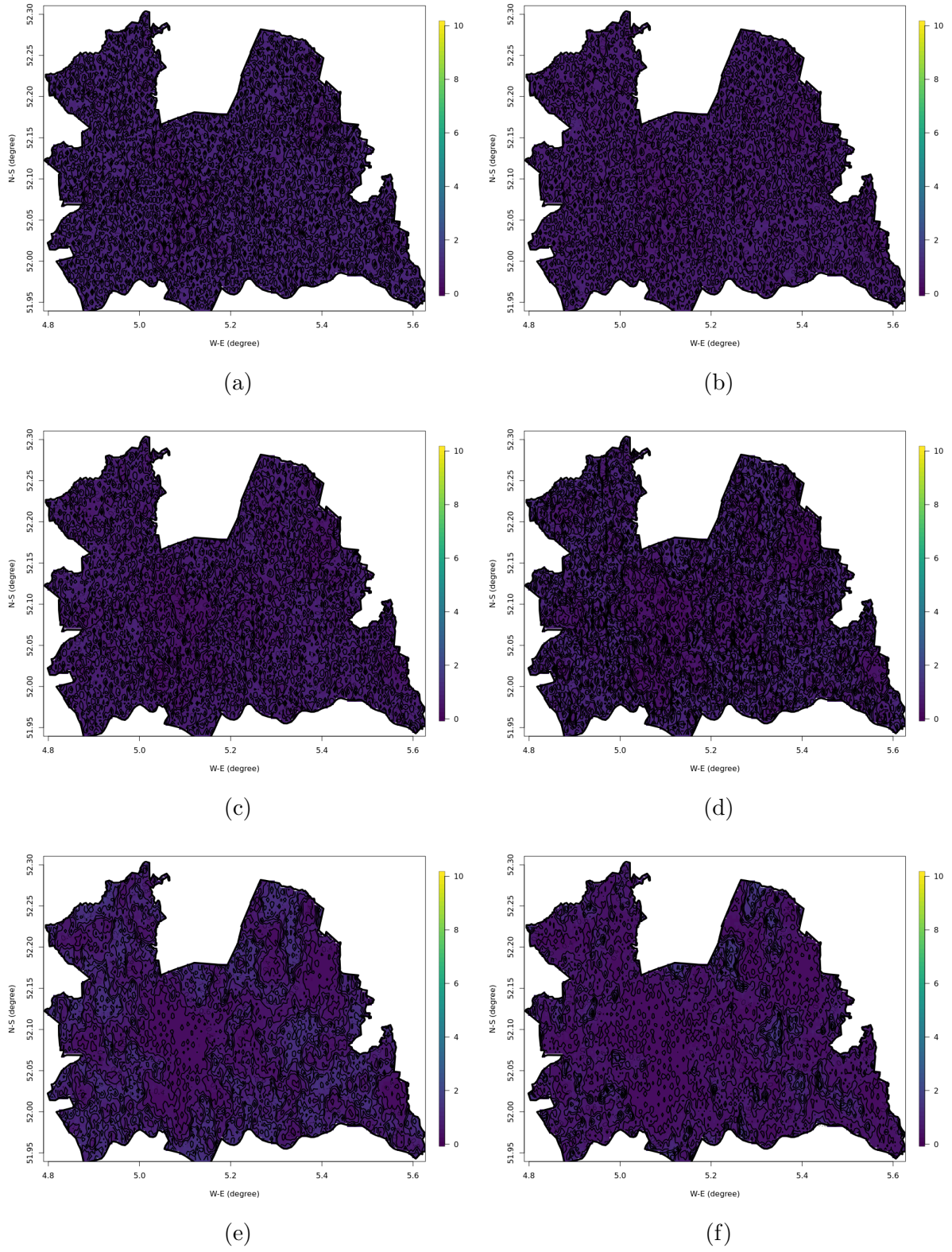



Figure A.9: Posterior SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with existing "top-down" estimates as normal priors and weighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

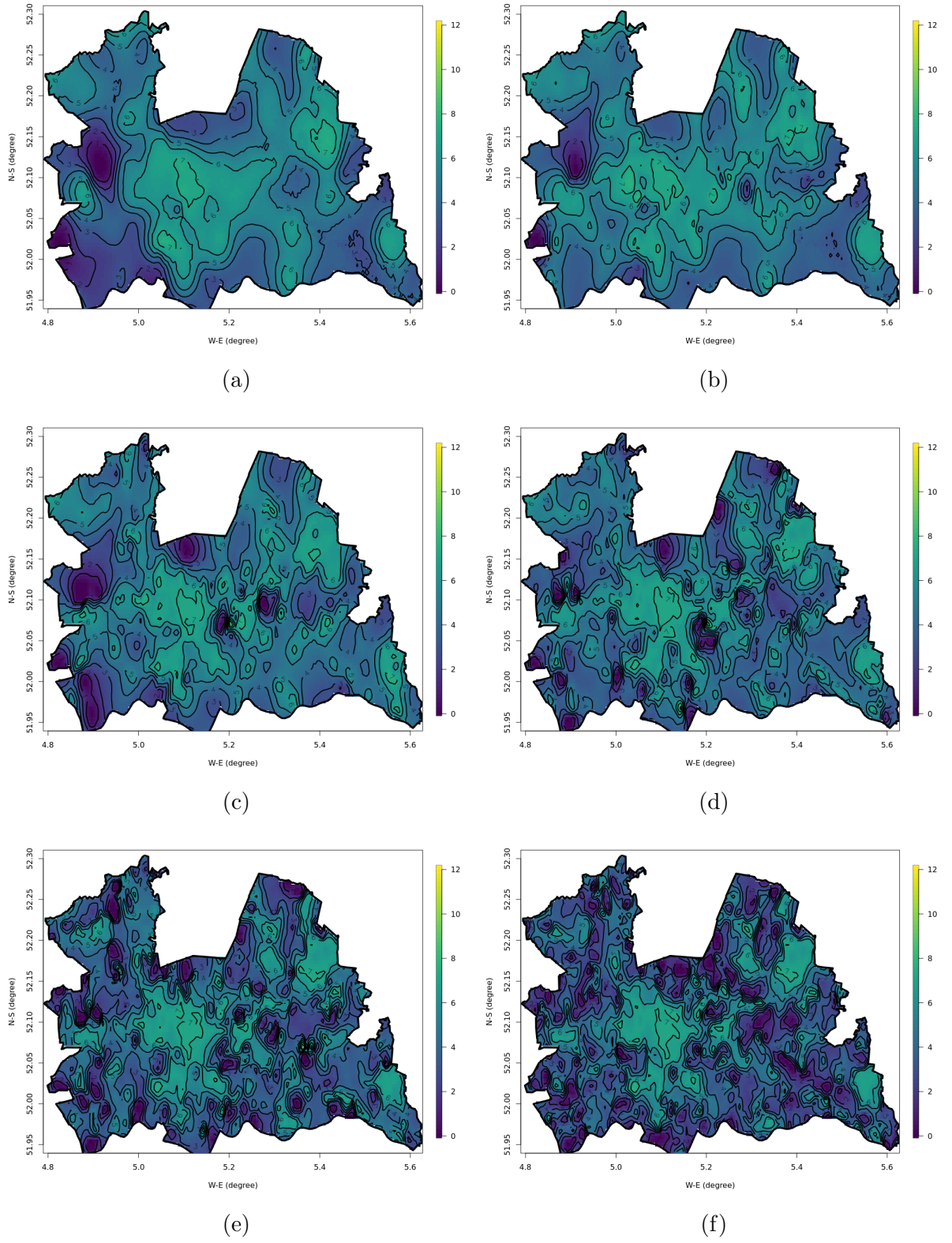



Figure A.10: Posterior mean of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with existing "top-down" estimates as PC priors and weighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

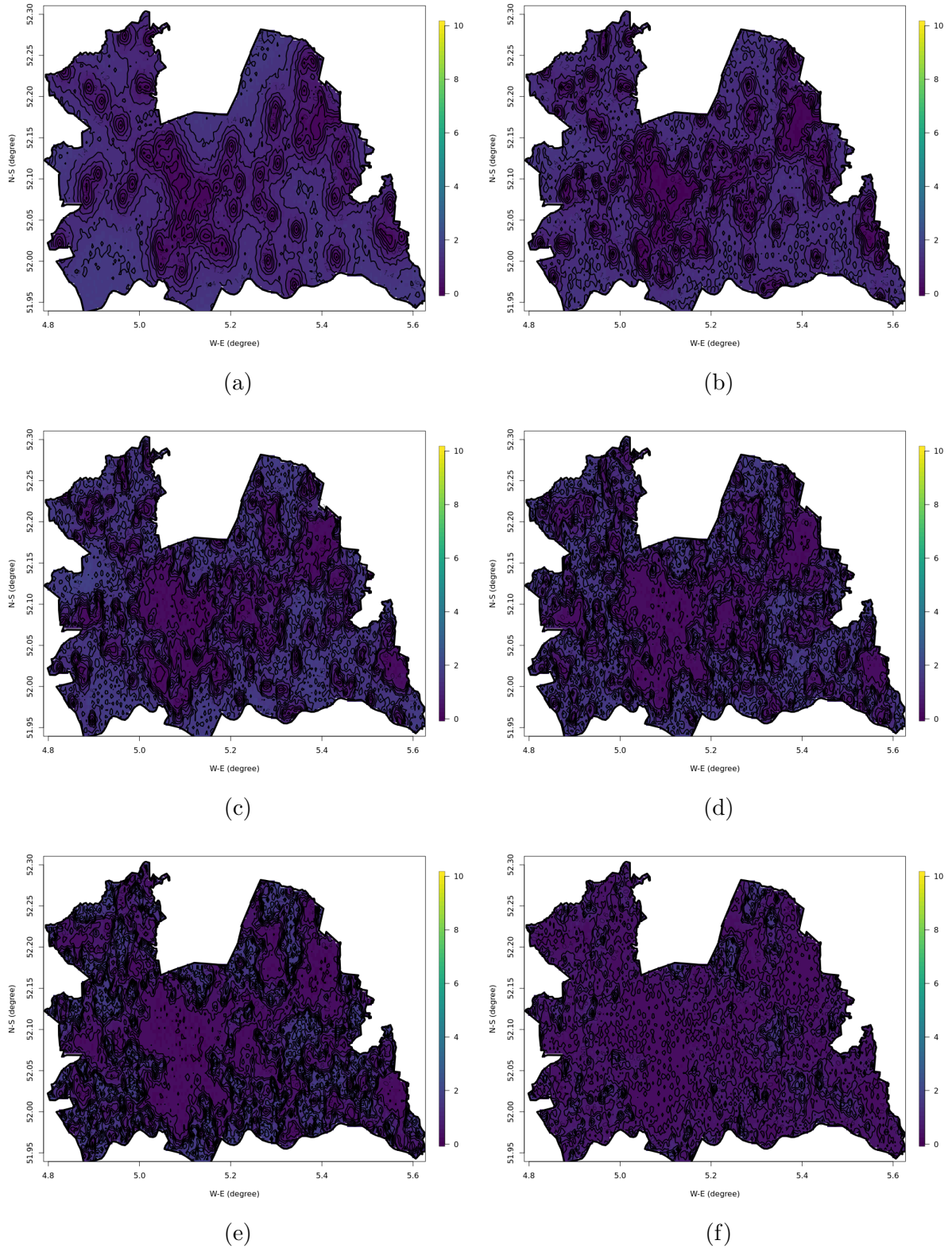



Figure A.11: Posterior SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with existing "top-down" estimates as PC priors and weighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

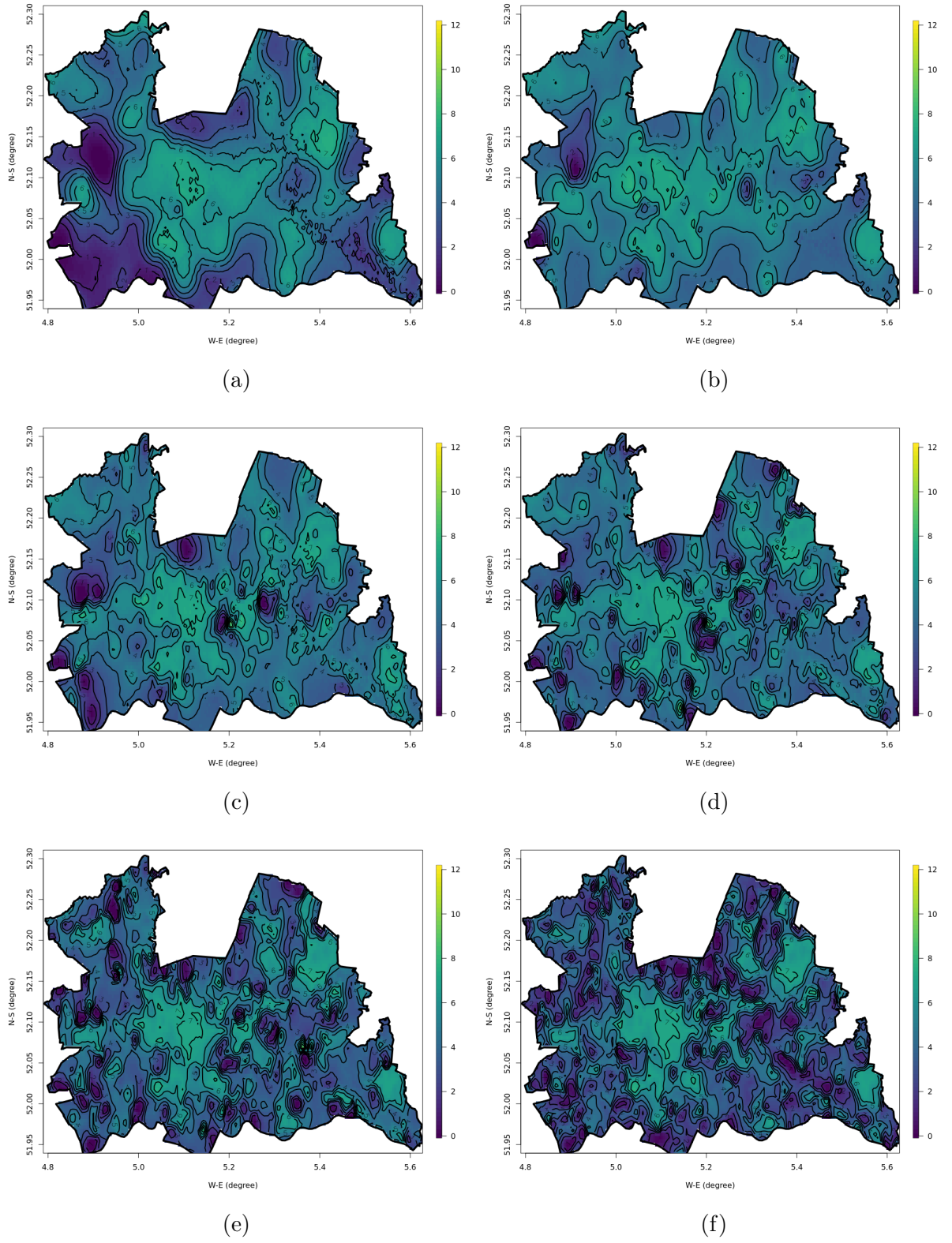



Figure A.12: Posterior mean of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with vague priors and weighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

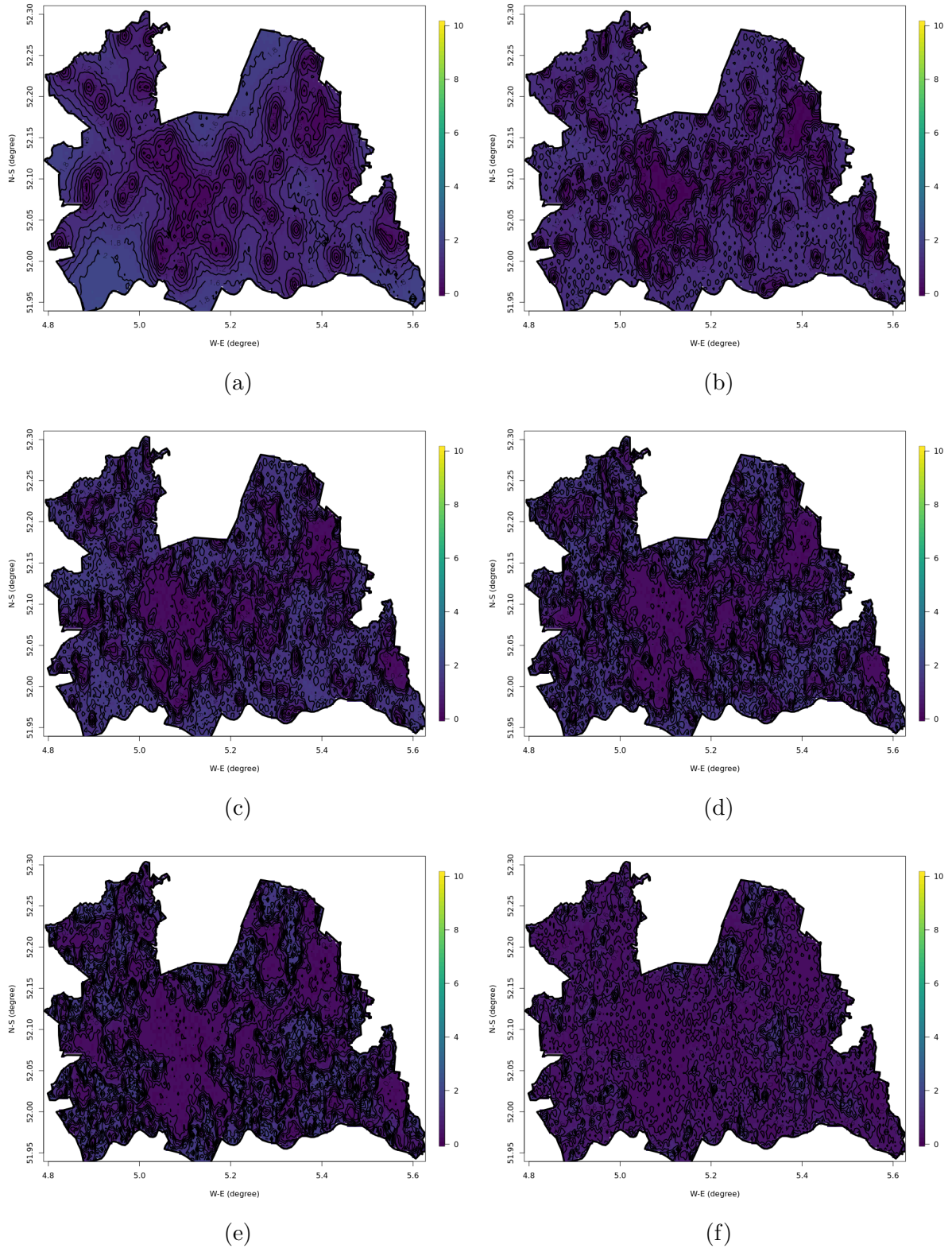



Figure A.13: Posterior SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with vague priors and weighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

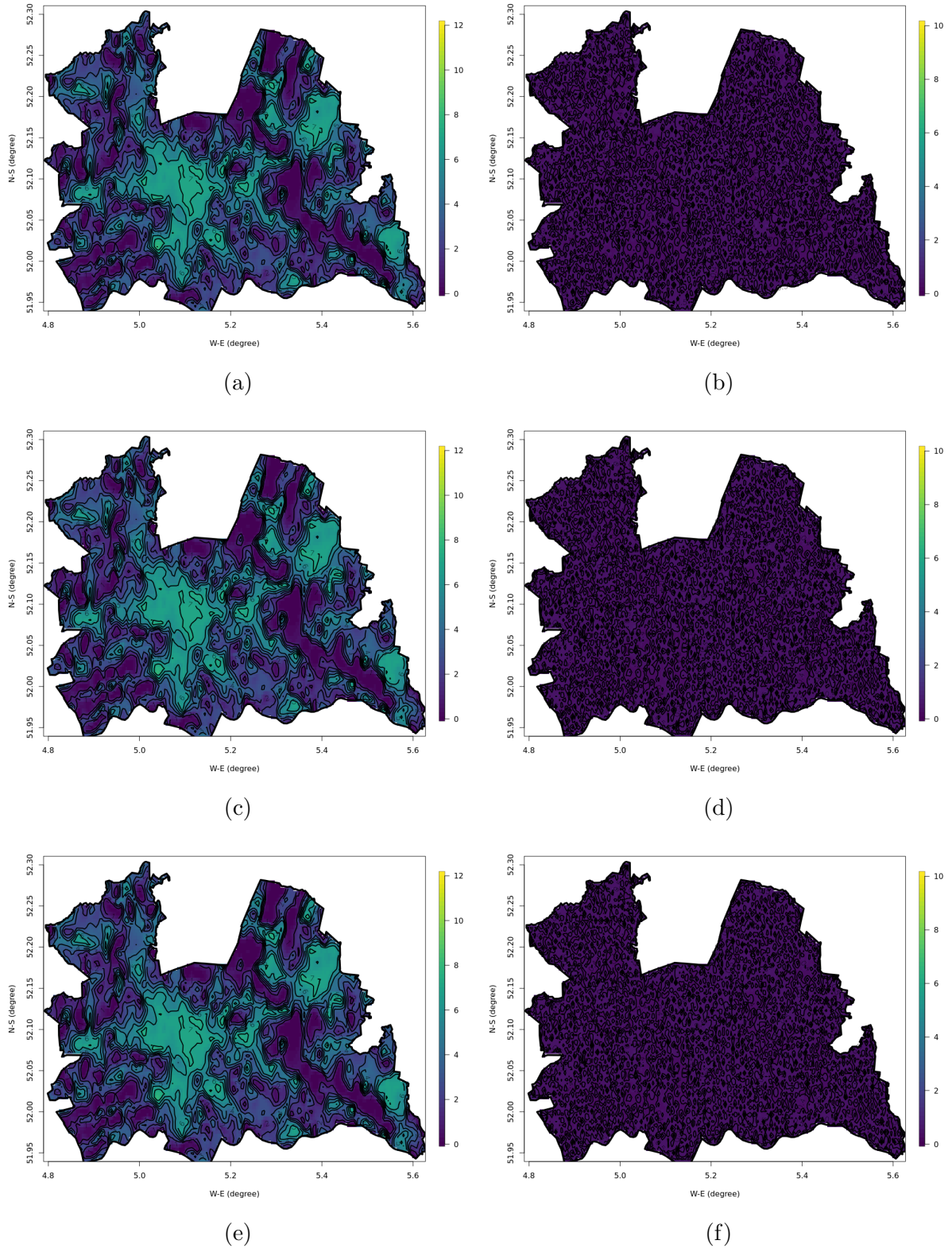
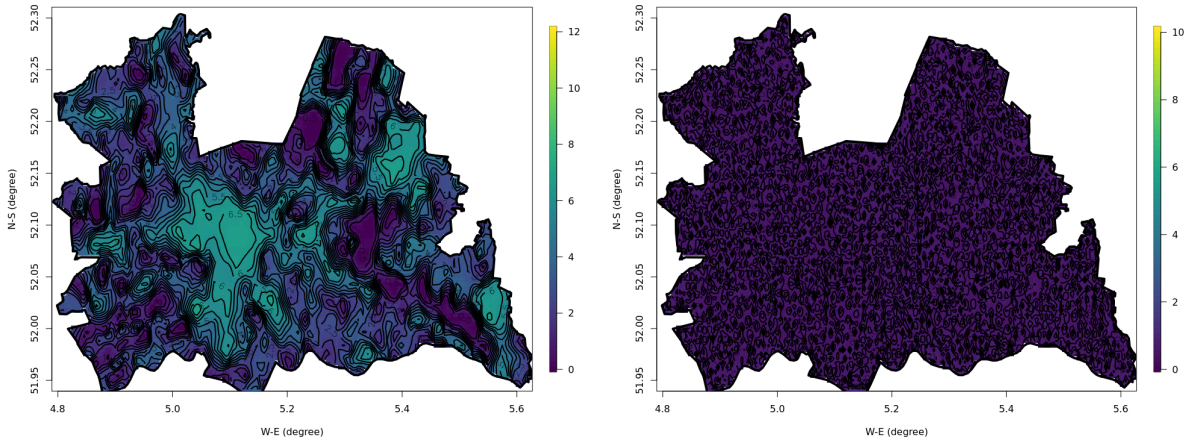



Figure A.14: Posterior mean and SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with (a, b) normal priors, (c, d) PC priors and (e, f) vague priors, and "bottom-up" samples of 100% real population data as the likelihoods.



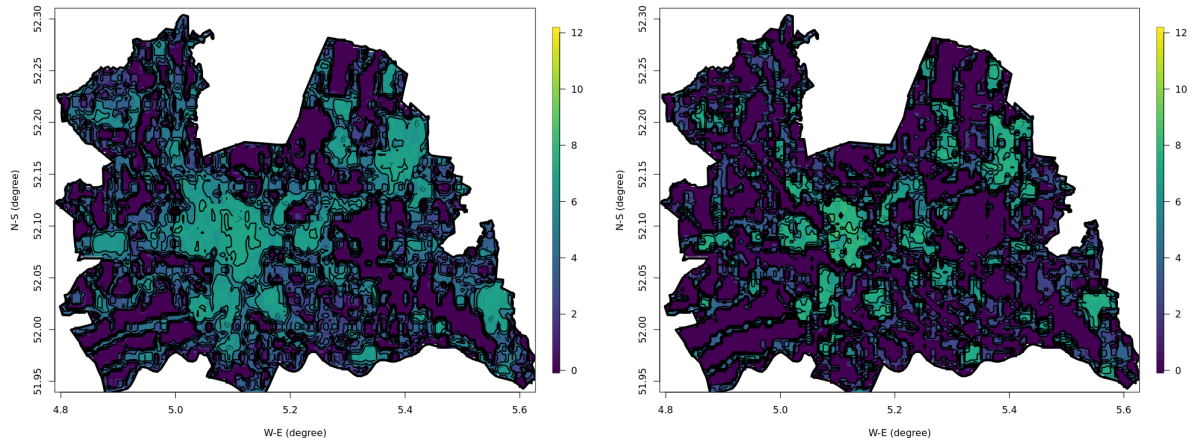

(a) (b)

Figure A.15: Posterior mean and SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations by projecting existing "top-down" estimates available originally at a resolution of 30 arc-seconds for free with vague priors.

# Appendix B

# Appendix for Chapter 5

The GHS-POP data at resolutions of 15 arc-seconds (i.e., the real population data in Chapter 4) and the CBS population data at the same resolution (i.e., the real population data in Chapter 5) are visualised as Figure B.1, with the transformation function $g_3(y_i)$ applied.




(a) (b)

Figure B.1: The (a) GHS-POP data at a resolution of 15 arc-seconds (i.e., the real population data defined in Chapter 4) and (b) CBS population data at the same resolution on the transformed scale $g_3(y_i)$.

Figure B.2 and Figure B.3, Figure B.4 and Figure B.5, Figure B.6 and Figure B.7, Figure B.8 and Figure B.9, Figure B.10 and Figure B.5, and Figure B.12 and Figure B.13 show the mapped posterior means and SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of the thirty-six estimations made with (i) unweighted sampling and normal priors, (ii) unweighted sampling and PC priors, (iii) unweighted sampling and vague priors, (iv) weighted sampling and normal priors, (v) weighted sampling and PC priors, and (vi) weighted sampling and vague priors, and different sample sizes, comparable to the map of the real population data defined in this chapter and shown as Figure B.1b.

Figure B.14 shows the mapped posterior means and SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of the estimations made with "bottom-up" sample of 100% full real population data.

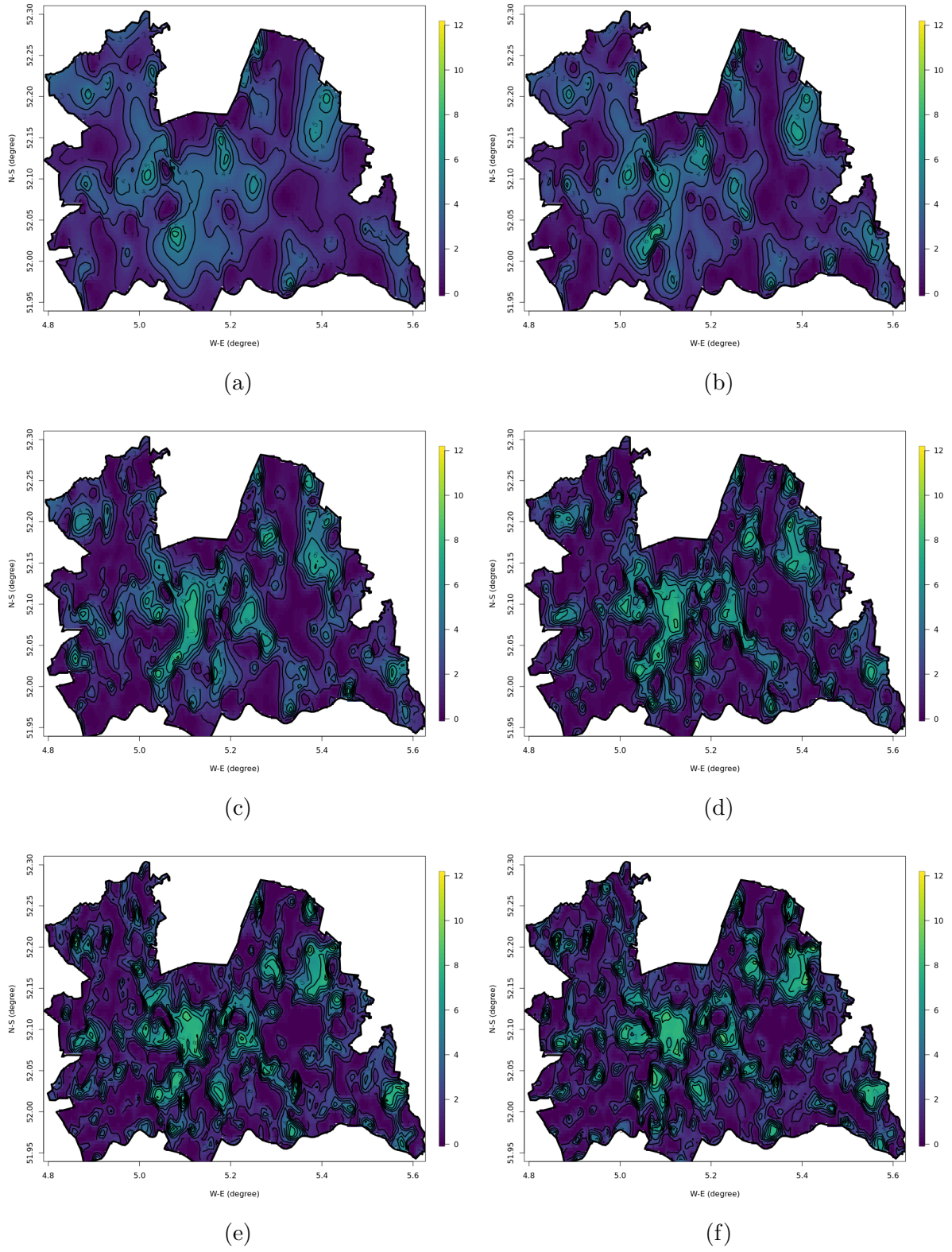



Figure B.2: Posterior mean of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with existing "top-down" estimates as normal priors and unweighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

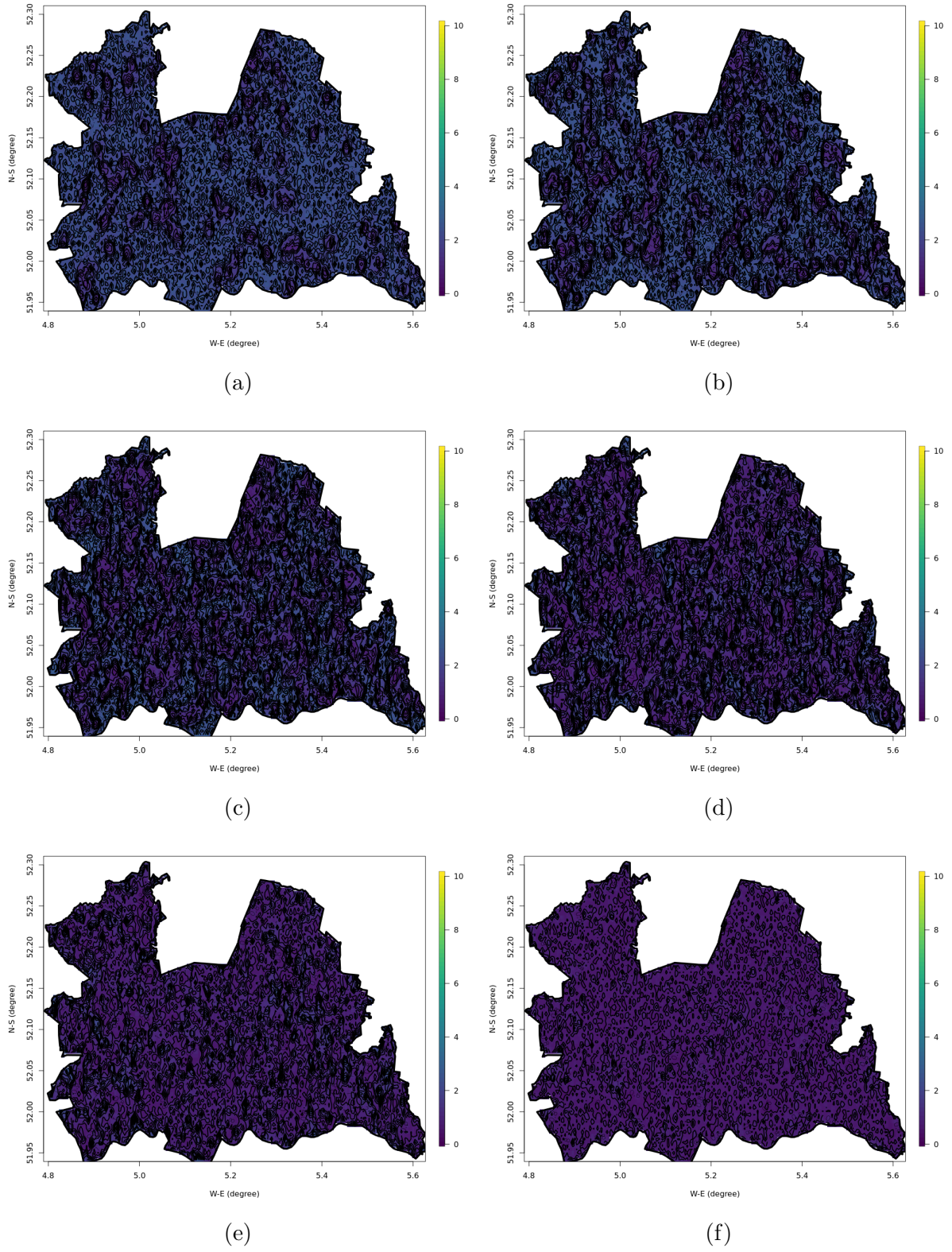



Figure B.3: Posterior SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with existing "top-down" estimates as normal priors and unweighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

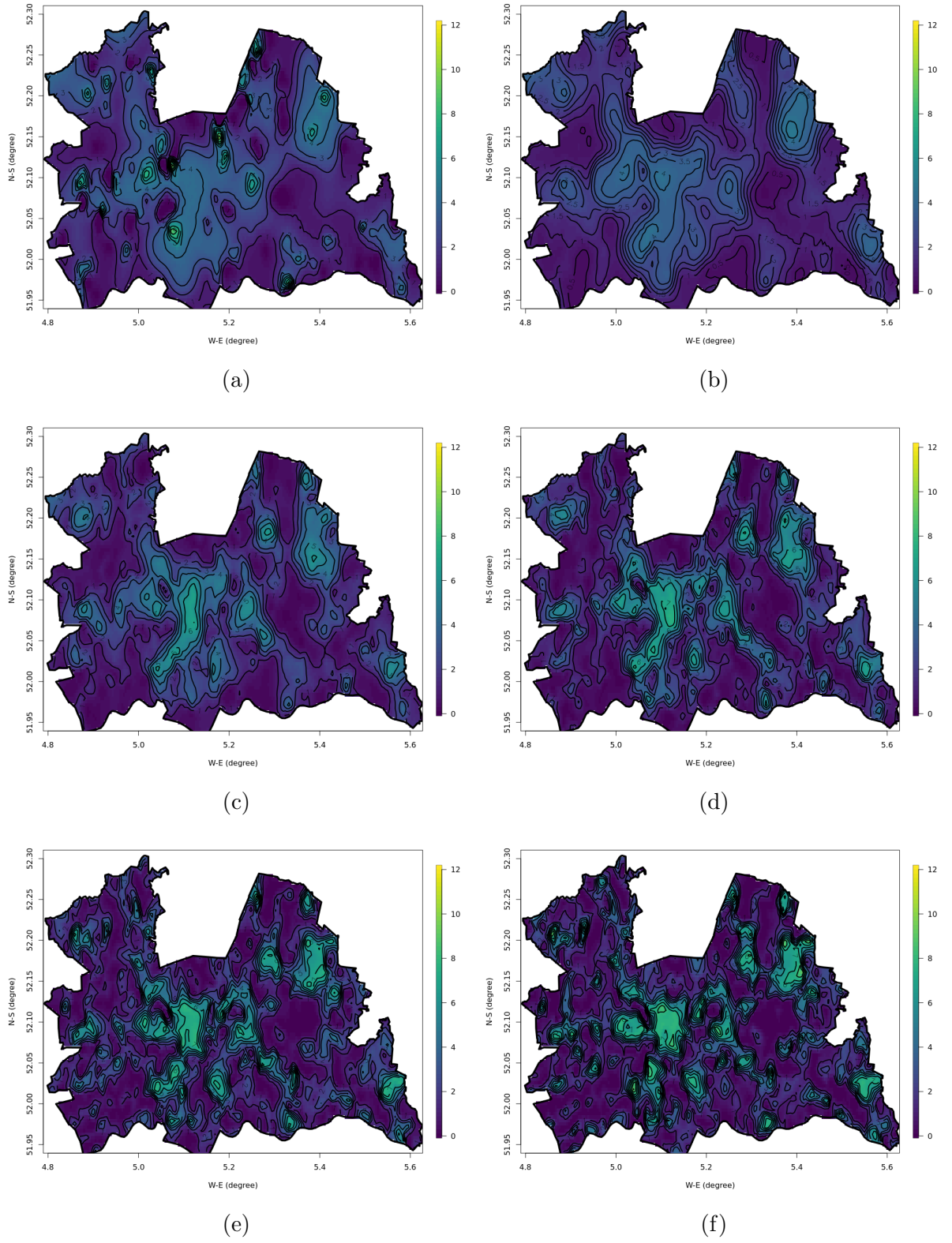



Figure B.4: Posterior mean of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with existing "top-down" estimates as PC priors and unweighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

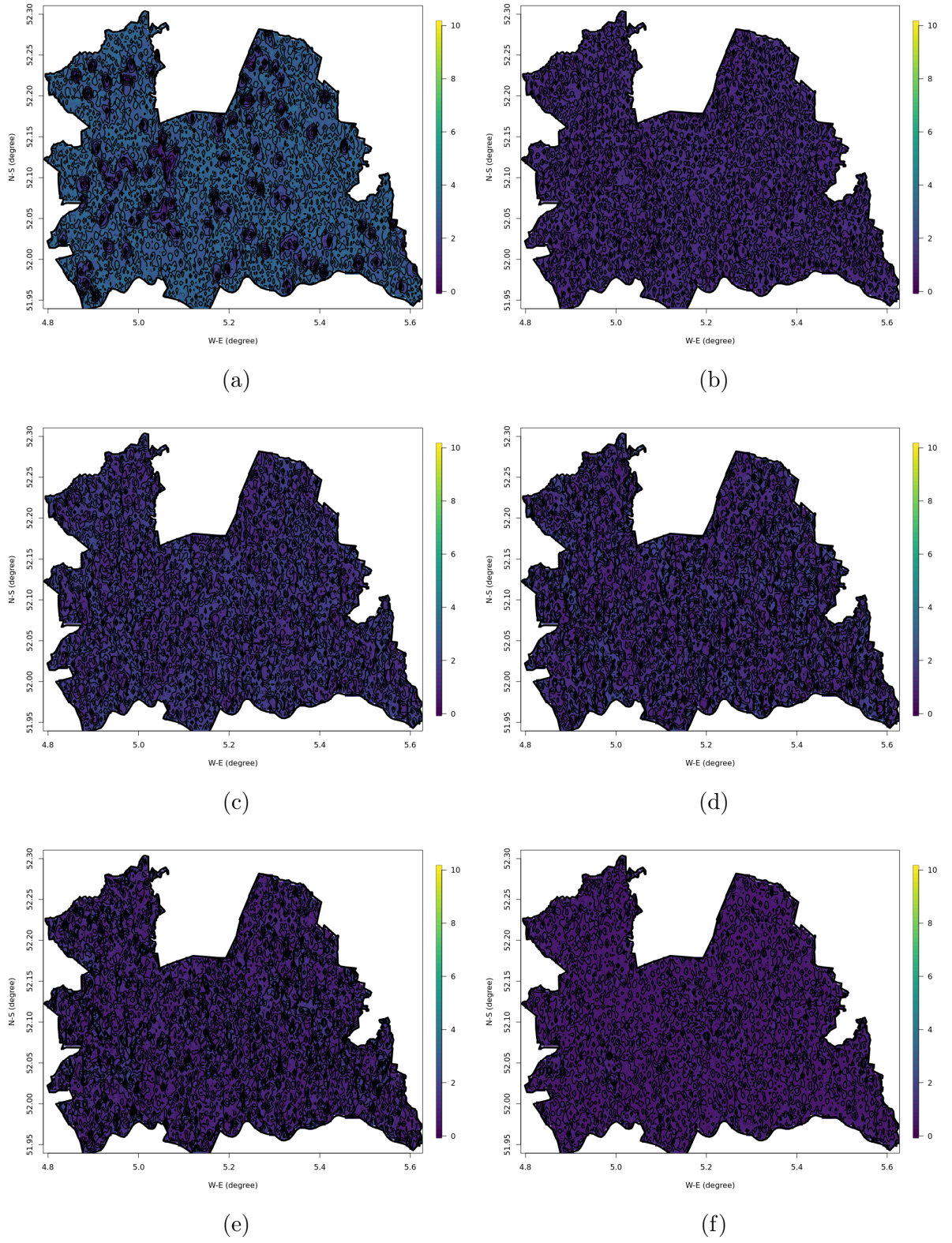



Figure B.5: Posterior SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with existing "top-down" estimates as PC priors and unweighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

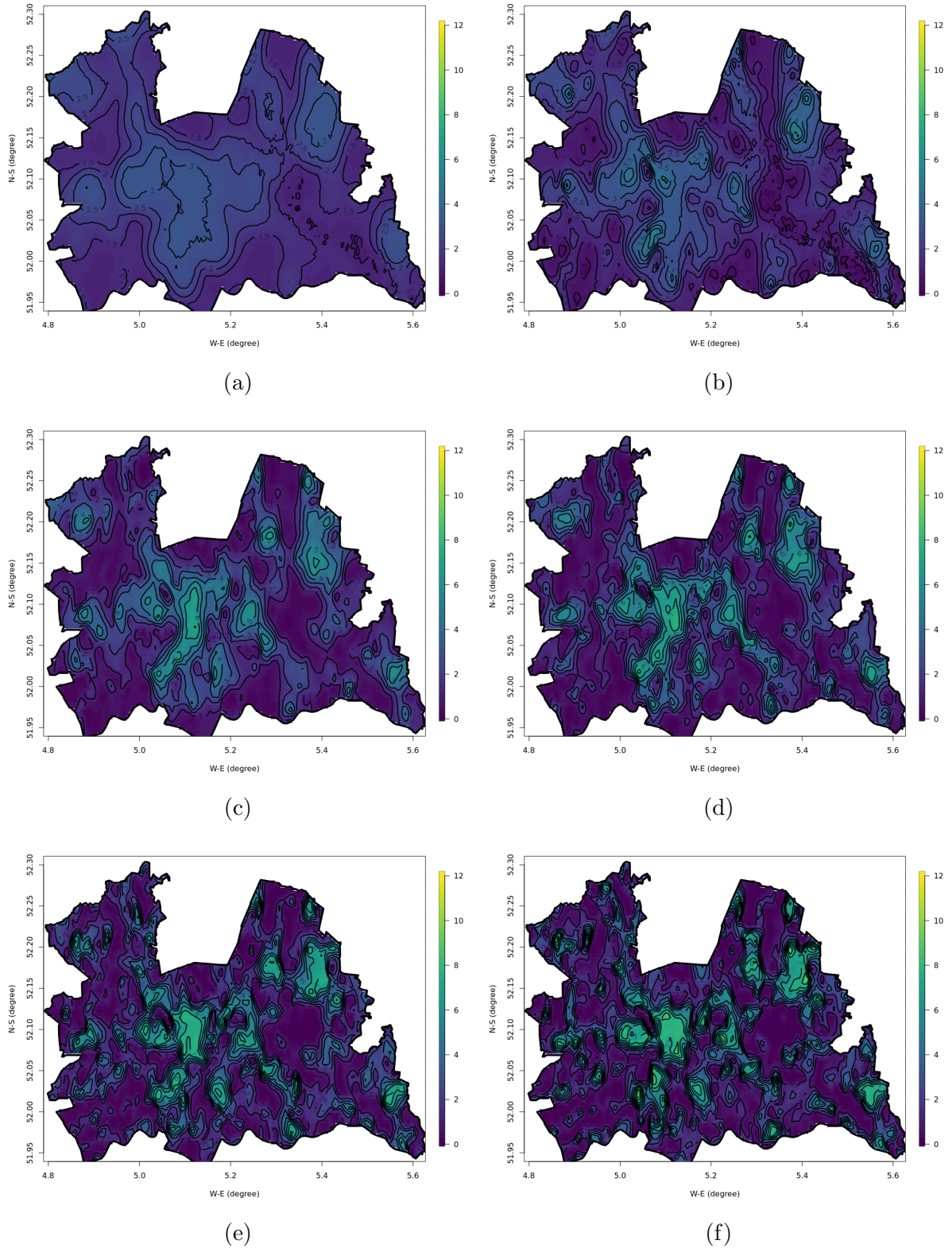



Figure B.6: Posterior mean of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with vague priors and unweighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

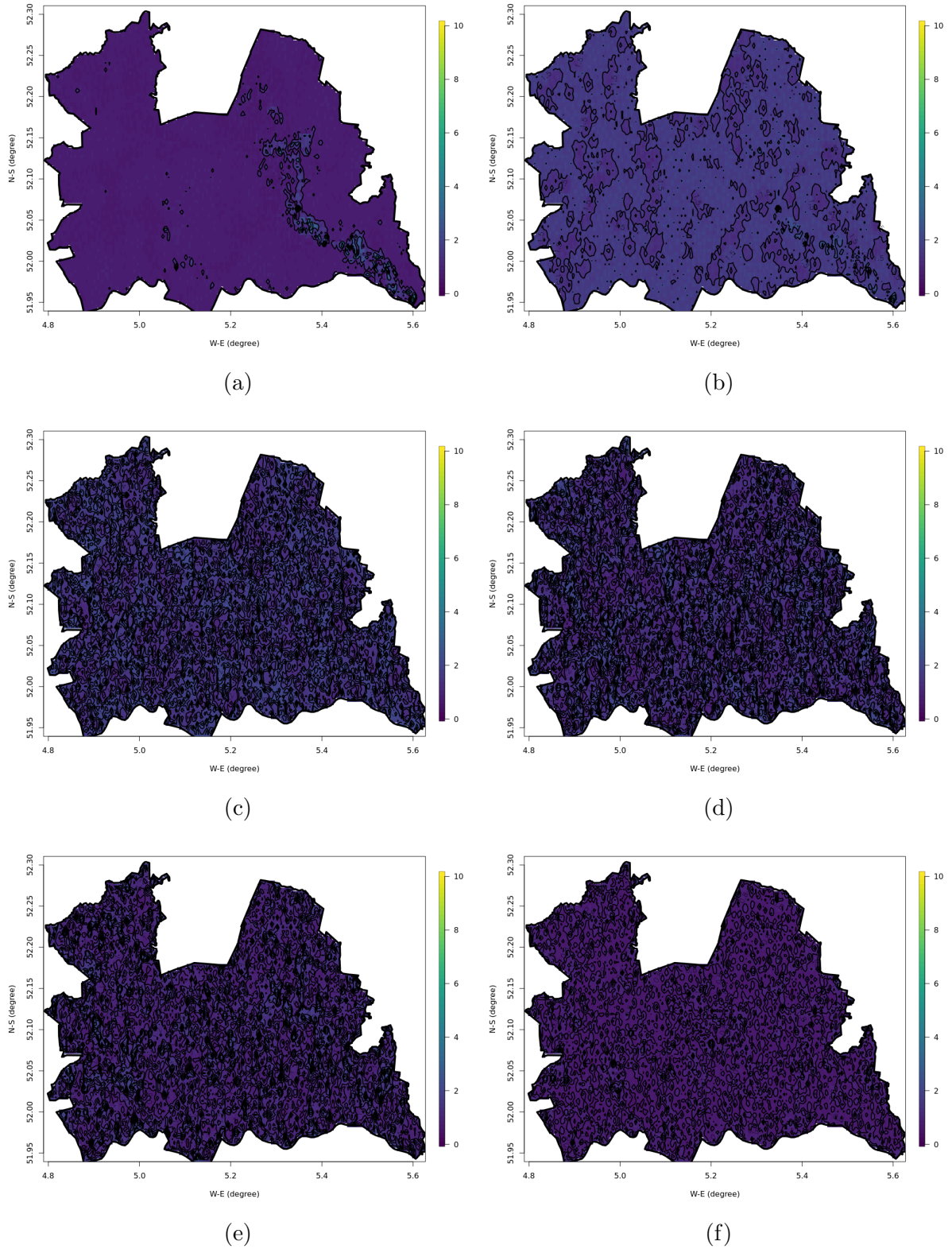



Figure B.7: Posterior SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with vague priors and unweighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

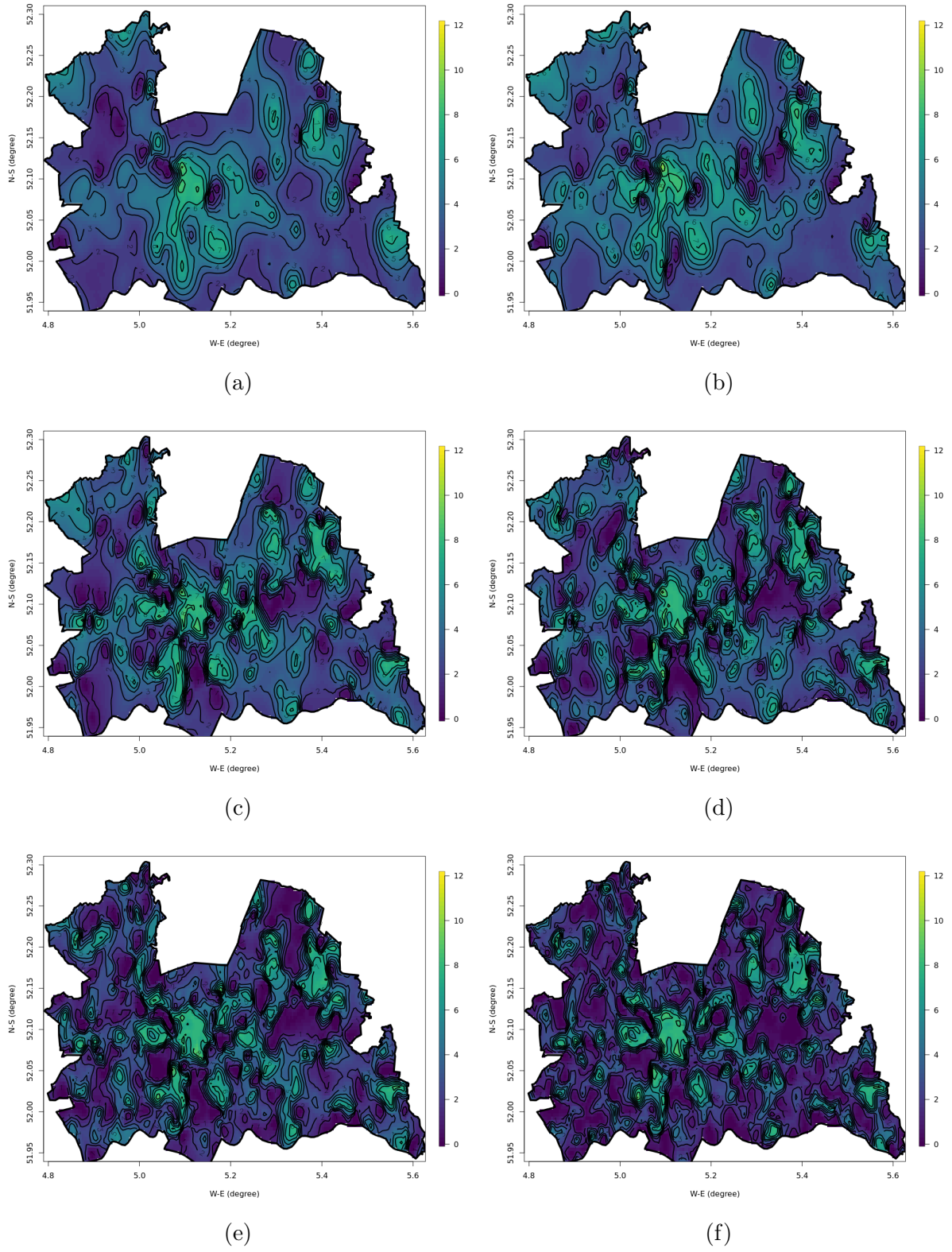



Figure B.8: Posterior mean of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with existing "top-down" estimates as normal priors and weighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

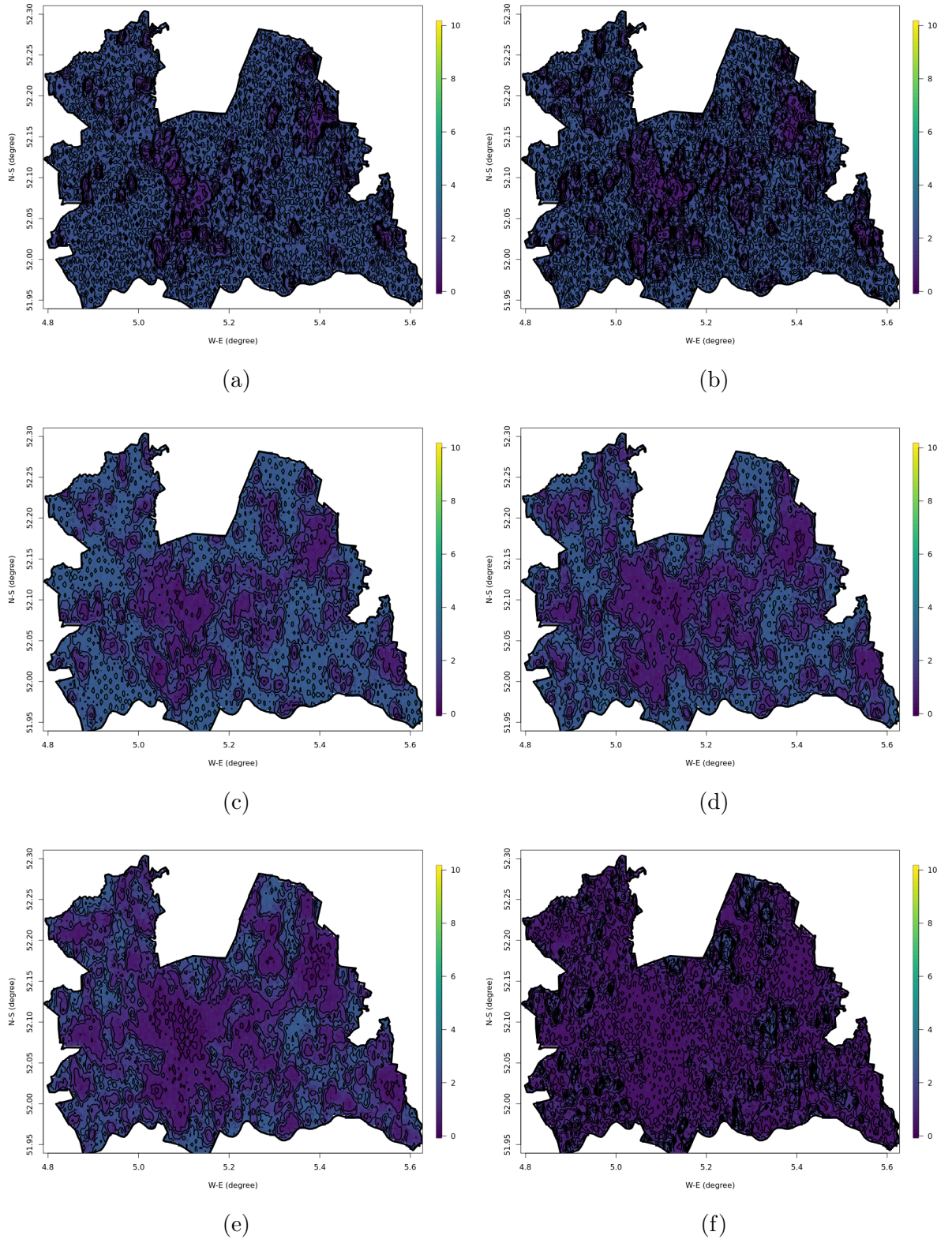



Figure B.9: Posterior SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with existing "top-down" estimates as normal priors and weighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

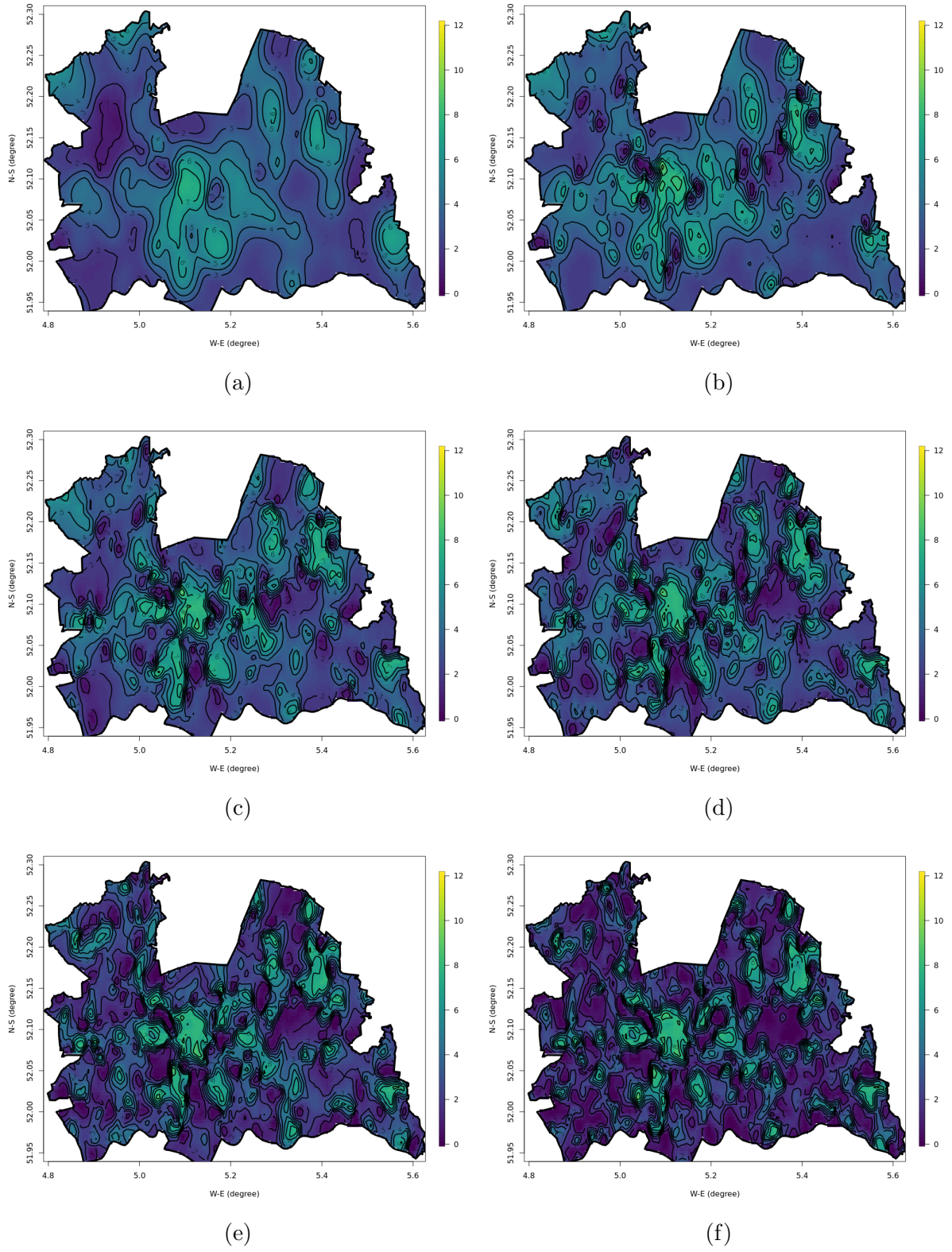



Figure B.10: Posterior mean of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with existing "top-down" estimates as PC priors and weighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

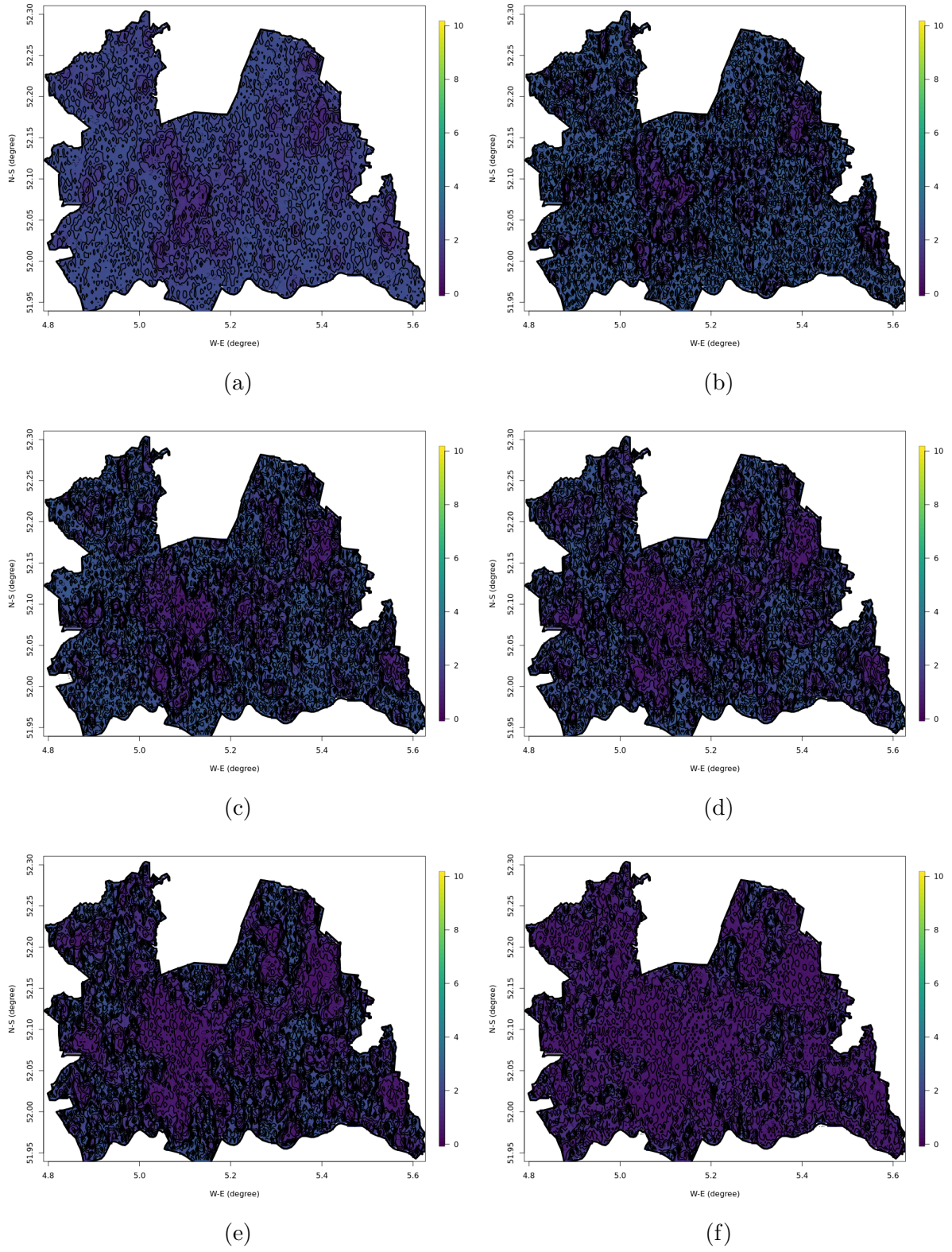



Figure B.11: Posterior SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with existing "top-down" estimates as PC priors and weighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

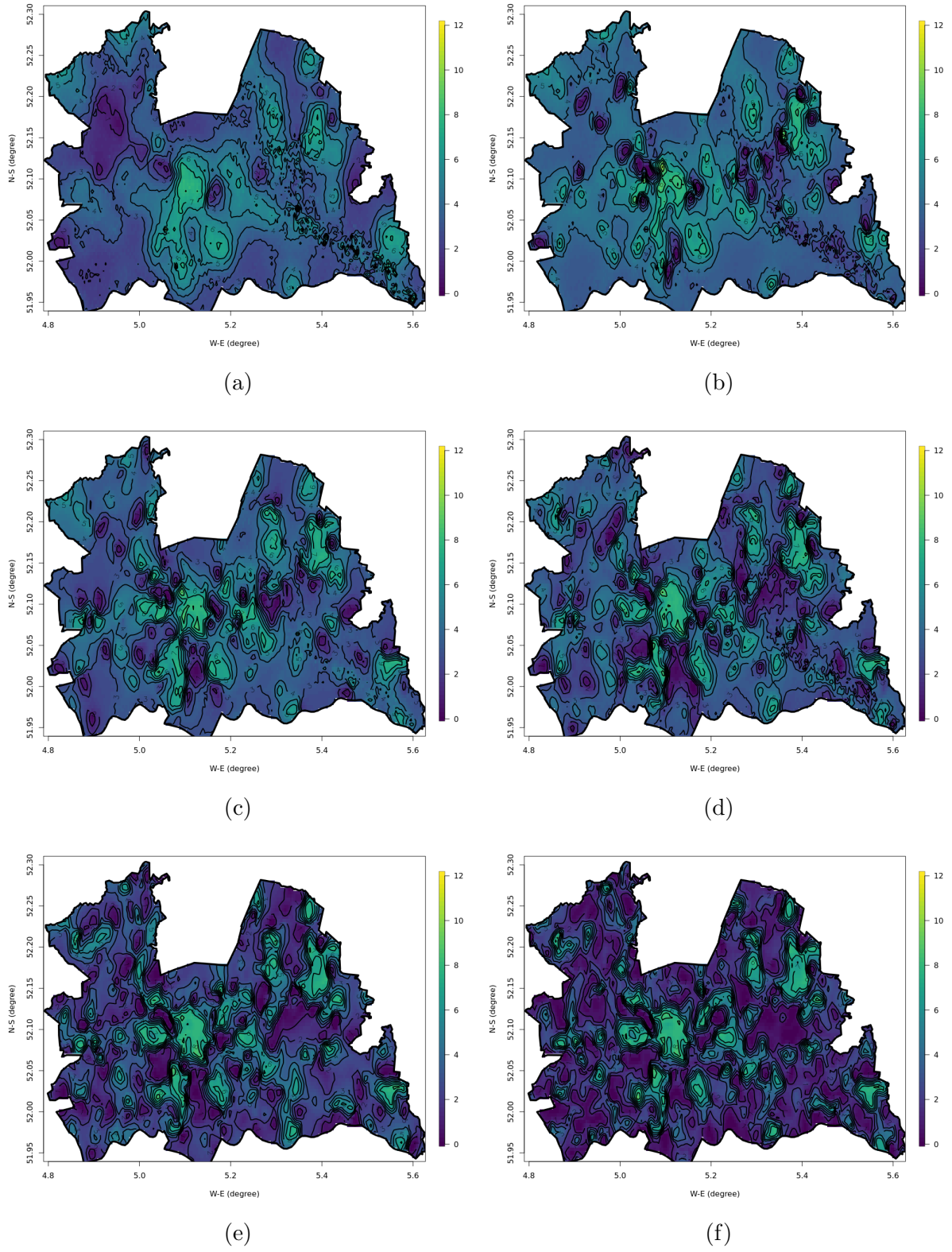



Figure B.12: Posterior mean of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with vague priors and weighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

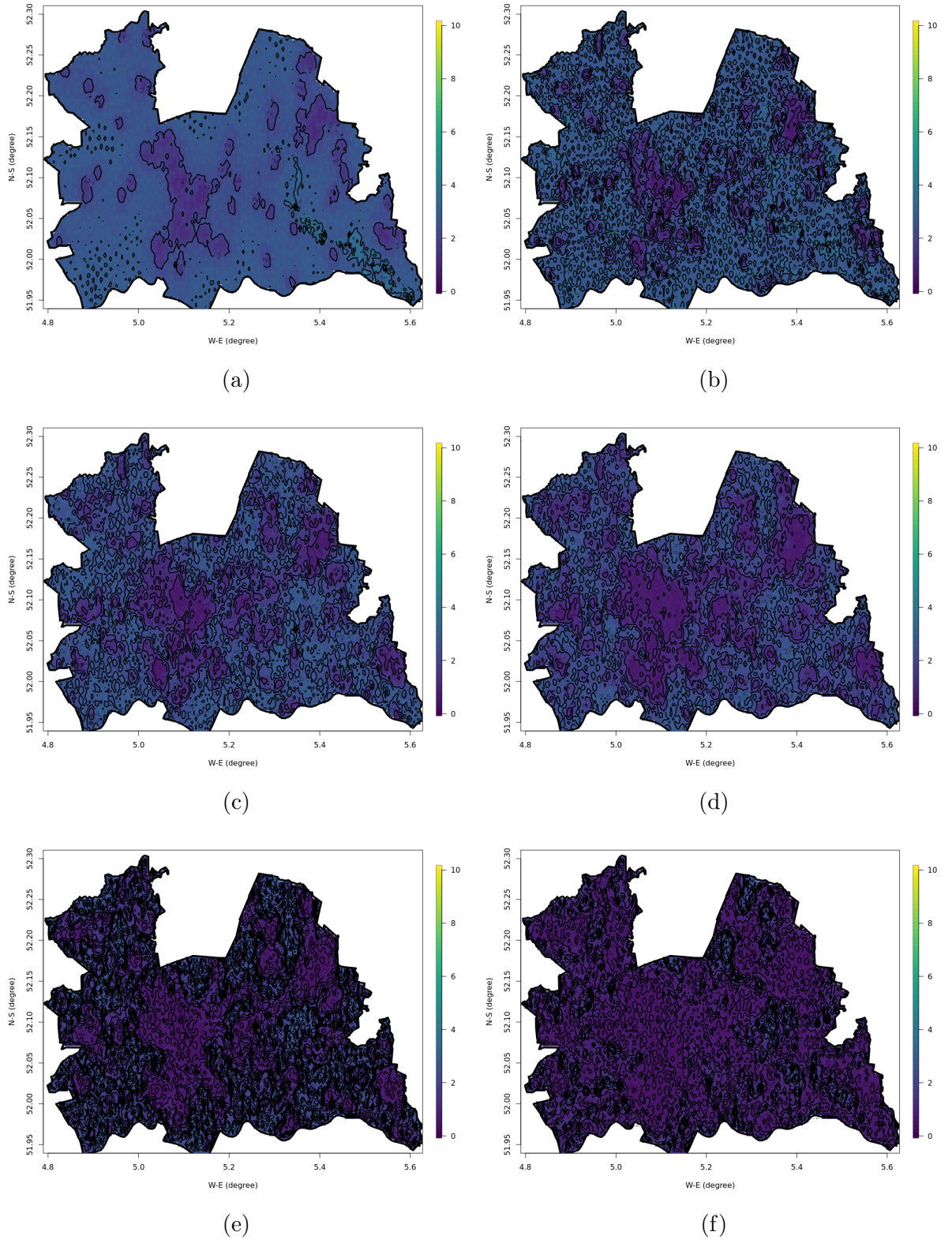



Figure B.13: Posterior SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with vague priors and weighted "bottom-up" samples of (a) 1%, (b) 2%, (c) 5%, (d) 10%, (e) 20% and (f) 50% real population data as the likelihoods.

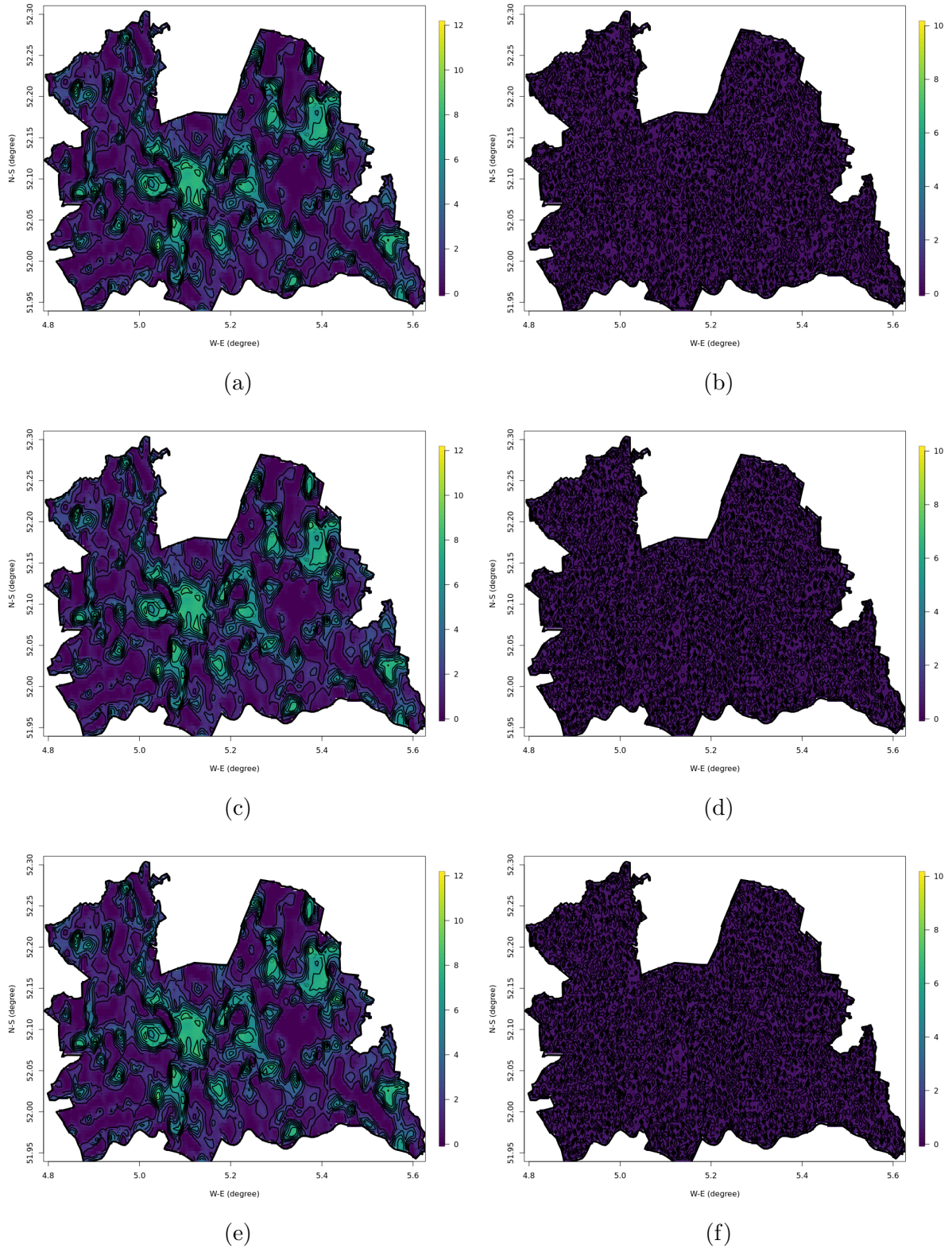



Figure B.14: Posterior mean and SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 15 arc-seconds on the basis of estimations made with (a, b) normal priors, (c, d) PC priors and (e, f) vague priors, and "bottom-up" samples of 100% real population data as the likelihoods.

# Appendix C

# Appendix for Chapter 6

The WorldPop data at a resolution of 30 arc-seconds for the years of 2002, 2005, 2008, 2011, 2014 and 2017 (i.e., Group 1) and the years of 2003, 2006, 2009, 2012, 2015 and 2018 (i.e., Group 2) are visualised as Figure C.1 and Figure C.2 respectively with the transformation function $g_3(y_i)$ applied.

Figure C.3, Figure C.4, Figure C.5, and Figure C.6 show the mapped posterior means and SD of the population counts on the transformed scale $g_3(y_i)$ at a resolution of 30 arc-seconds on the basis of the estimation made with all the WorldPop data in the estimation group for prediction on Group 1 and Group 2.

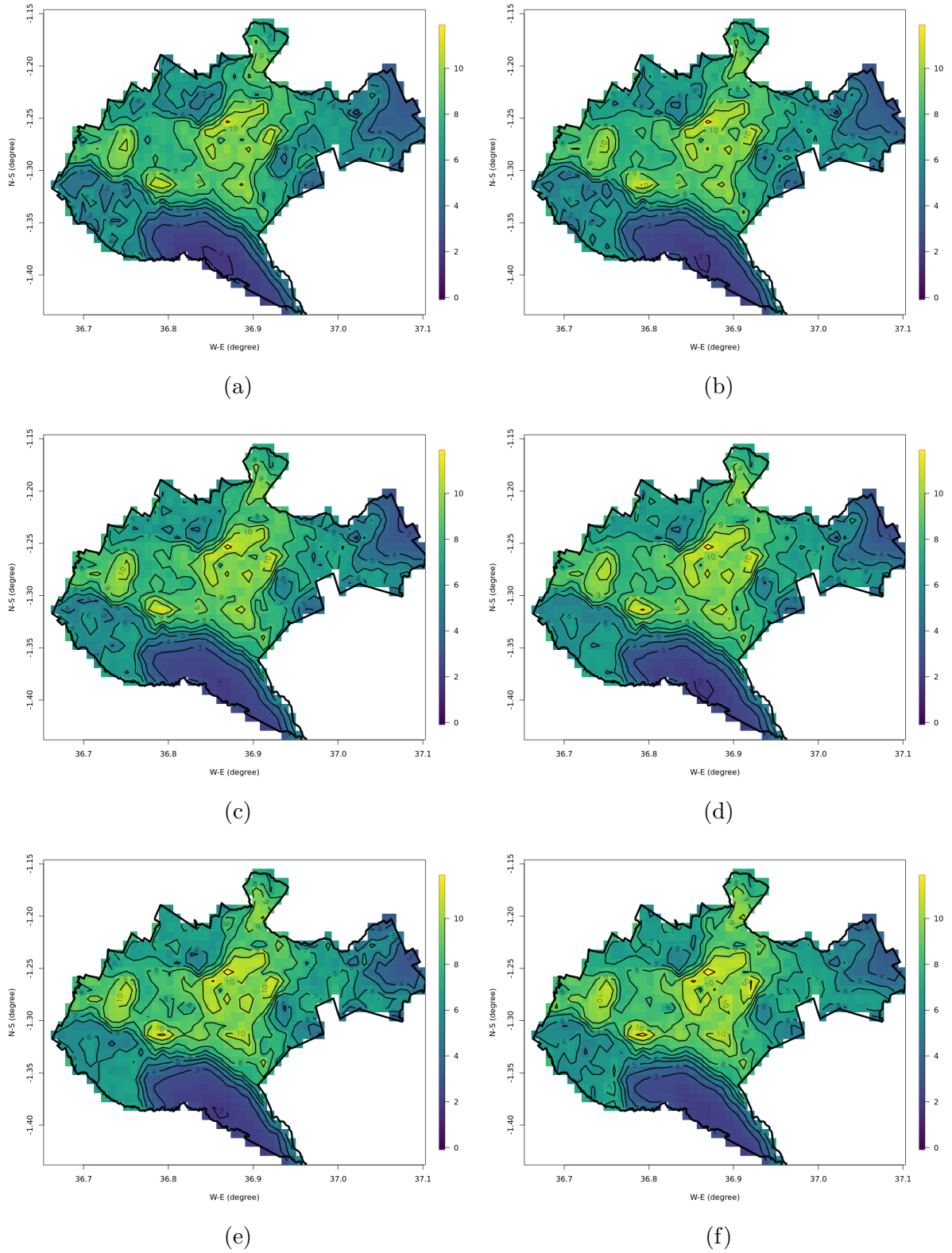



Figure C.1: The WorldPop data at a resolution of 30 arc-seconds for the years of (a) 2002, (b) 2005, (c) 2008, (d) 2011, (e) 2014 and (f) 2017 on the transformed scale $g_3(y_i)$ (i.e., Group 1).

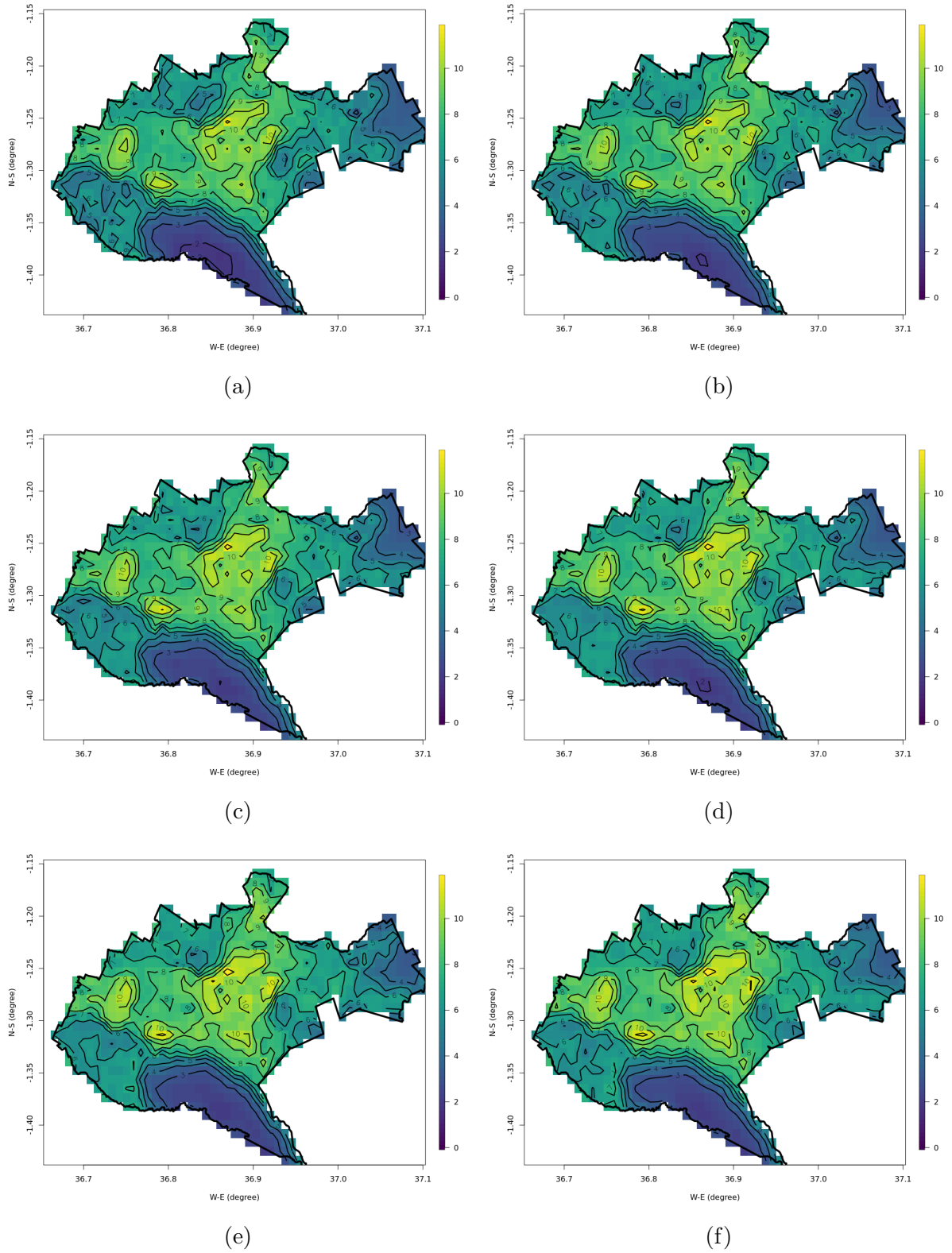



Figure C.2: The WorldPop data at a resolution of 30 arc-seconds for the years of (a) 2003, (b) 2006, (c) 2009, (d) 2012, (e) 2015 and (f) 2018 on the transformed scale $g_3(y_i)$ (i.e., Group 2).

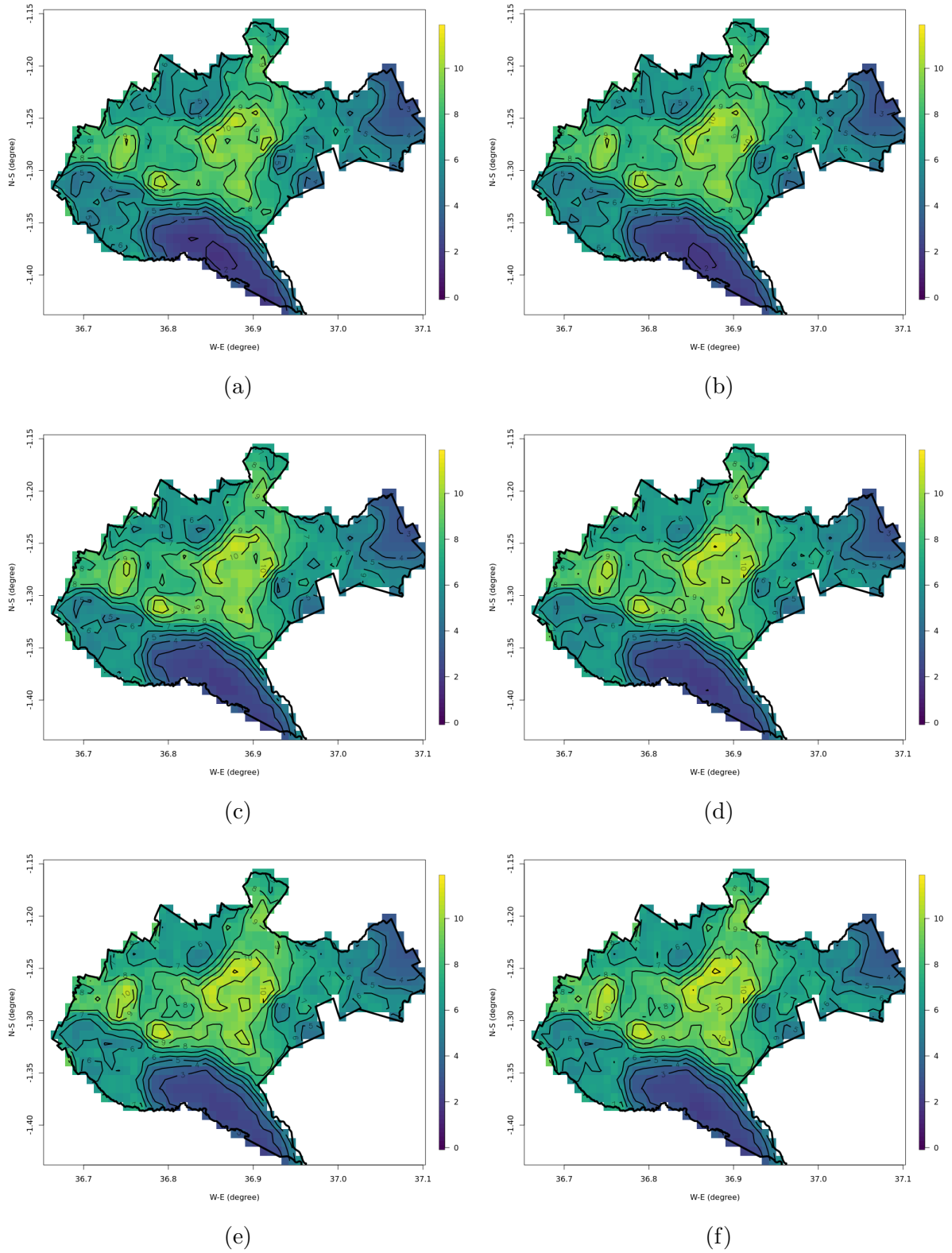



Figure C.3: Posterior mean of population counts on the transformed scale $g_3(y_i)$ at a resolution of 30 arc-seconds for the year of (a) 2002, (b) 2005, (c) 2008, (d) 2011, (e) 2014 and (f) 2017 for prediction on Group 1.

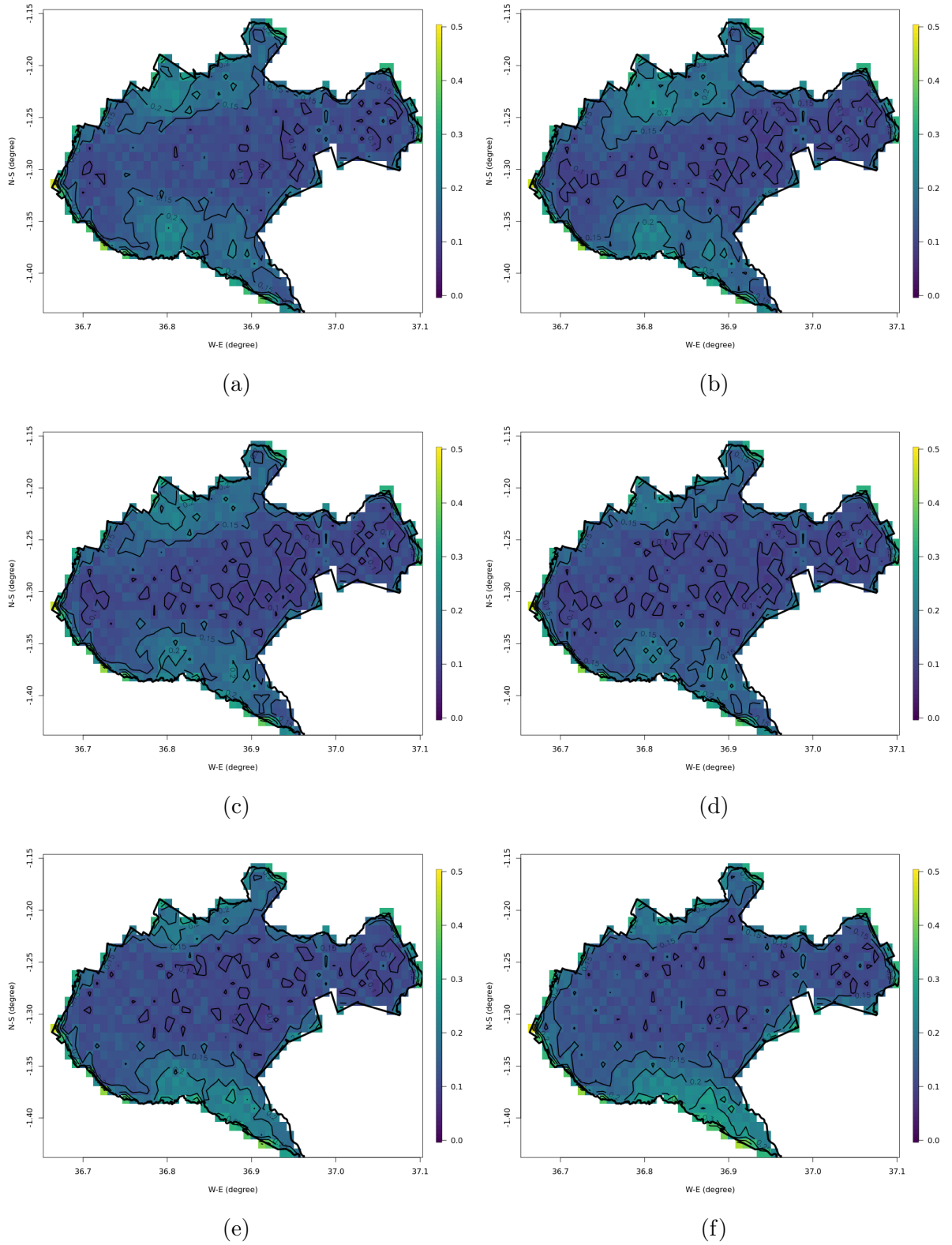



Figure C.4: Posterior SD of population counts on the transformed scale $g_3(y_i)$ at a resolution of 30 arc-seconds for the year of (a) 2002, (b) 2005, (c) 2008, (d) 2011, (e) 2014 and (f) 2017 for prediction on Group 1.

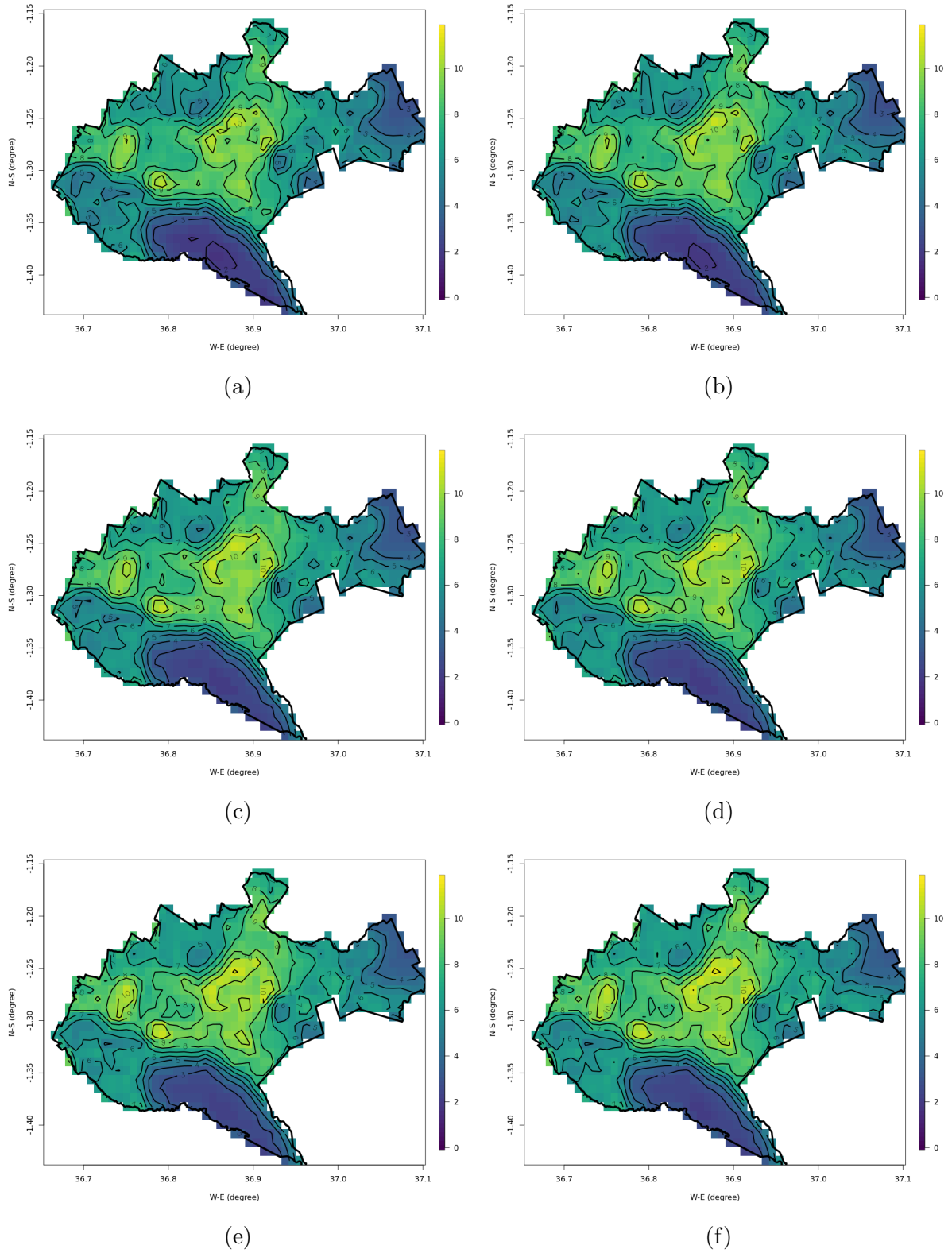



Figure C.5: Posterior mean of population counts on the transformed scale $g_3(y_i)$ at a resolution of 30 arc-seconds for the year of (a) 2003, (b) 2006, (c) 2009, (d) 2012, (e) 2015 and (f) 2018 for prediction on Group 2.

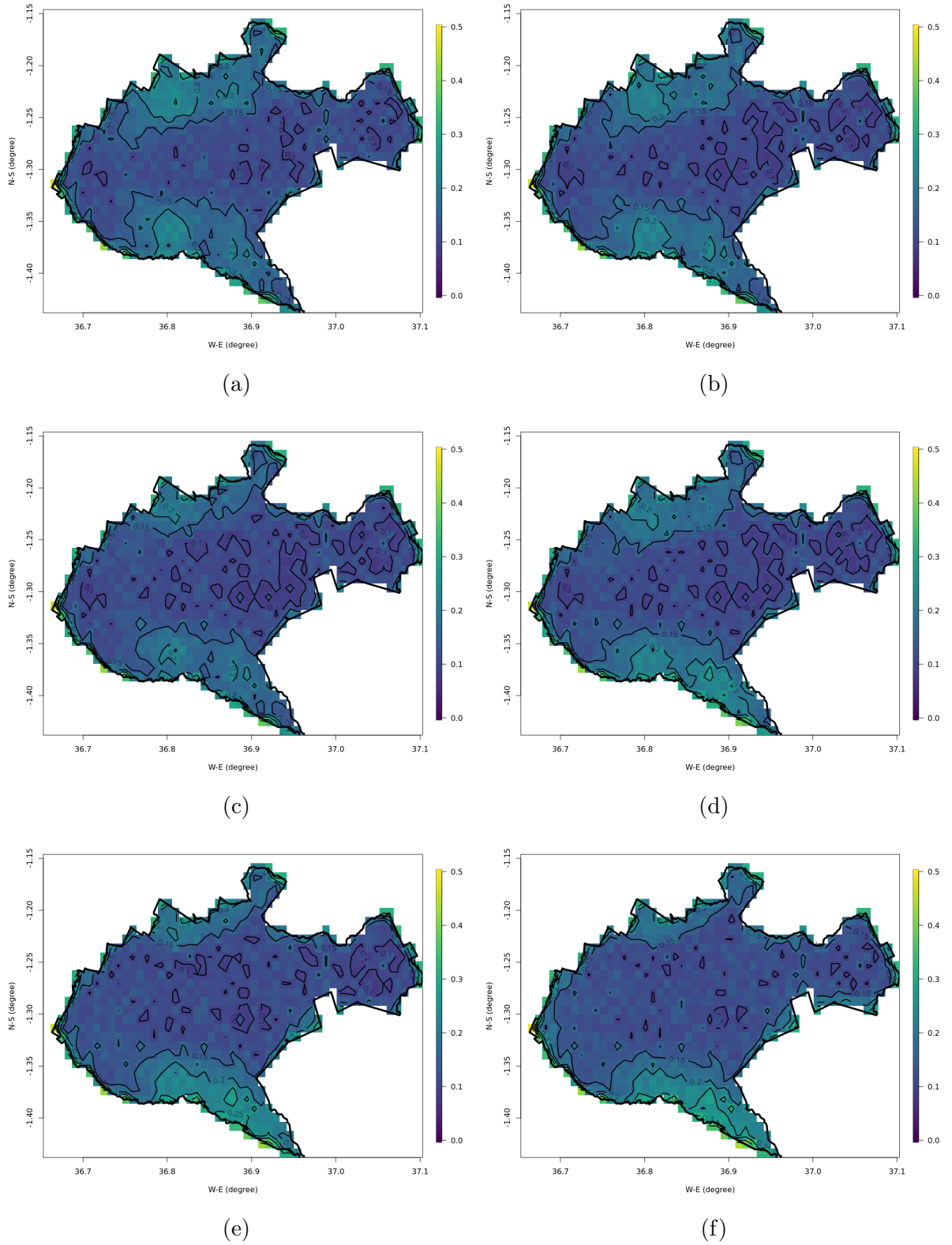



Figure C.6: Posterior SD of population counts on the transformed scale $g_3(y_i)$ at a resolution of 30 arc-seconds for the year of (a) 2003, (b) 2006, (c) 2009, (d) 2012, (e) 2015 and (f) 2018 for prediction on Group 2.

**LEUVEN STATISTICS RESEARCH CENTRE**
LStat, KU Leuven
Celestijnenlaan 200B Bus 5307
B-3001 Leuven (Heverlee), BELGIË
tel. +32 16 32 88 75
lstat.kuleuven.be