Skewed Spatial Modeling for Arsenic Contamination in Bangladesh

Qi Zhang

A Thesis
in
The Department
of
Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Arts (Mathematics) at
Concordia University
Montréal, Québec, Canada

July 2019

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Qi Zhang

Entitled: Skewed Spatial Modeling for Arsenic Contamination in Bangladesh

and submitted in partial fulfillment of the requirements for the degree of

Master of Arts (Mathematics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

_____ Examiner

Dr. Arusharka Sen

_____ Examiner

Dr. Frédéric Godin

_____ Thesis Supervisor

Dr. Yogendra P. Chaubey

_____ Thesis Supervisor

Dr. Alexandra M. Schmidt

Approved by _____

Chair of Department or Graduate Program Director

_____

Dean of Faculty

Date _____

Abstract

Skewed Spatial Modeling for Arsenic Contamination in Bangladesh
by Qi Zhang

Bangladesh has been facing serious problem in arsenic contamination for more than two decades. Drinking and irrigating contaminated water put the health of more than 85 million people at risk. The project "*Groundwater Studies for Arsenic Contamination in Bangladesh*" led by British Geological Survey had been conducted during 1998 to 2001. A few studies have been carried out from different perspectives. The district Comilla is considered to be the most severed region with highest arsenic-related deaths according to Flanagan, Johnston, and Zheng, 2012.

In this thesis, we examine the arsenic groundwater concentration in Comilla district. We propose spatial models for making inference under Bayesian framework. We demonstrate that models based on the gamma distribution with spatial structure capture the characteristics of arsenic levels, appropriately compared to other models. We also perform spatial interpolation (kriging) to describe the situation of the arsenic levels across all Comilla.

Acknowledgements

I would like to express my deepest appreciation to my supervisors Dr. Yogendra P. Chaubey and Dr. Alexandra M. Schmidt. Dr. Chaubey has been giving me an incredible guidance and support through out my entire academic career in Canada, he always being patient, knowledgeable and wise. It is a great honor to work with Dr. Schmidt on my thesis, her great patience in guidance, depth in expertise and the attitude towards scientific research has been a role model to me. Working under such atmosphere make me feel nothing but pure joy during the research.

I would like to extend my sincere thanks to the committee Dr. Arusharka Sen and Dr. Frédéric Godin taking their valuable time reading and commenting my work.

在此我要特别感谢培养我的父亲张礼佳和母亲赵凌雪，感谢你们长期以来对我的关心、理解和支持，没有你们就不会有我今天的成绩。(Acknowledgement to my family)

Special thanks to all great professors in the department for your great lectures, especially to Dr. Xiaowen Zhou and Dr. Wei Sun who have been sharing their precious knowledge and experience to me over the years.

Appreciation should also go to all my friends and colleagues here for sharing an amazing life together. I enjoy every bit of moment on discussing math and lives.

Finally, I'm extremely grateful to Liverpool FC, the club I will support till the last second of my life. Thank you very much for showing me great spirit, pure passion and countless miracles consistently and many congratulations for winning the UEFA Champions League for the sixth time as well as the UEFA Super Cup for the fourth time. You'll Never Walk Alone.

'This material was produced by the British Geological Survey and the Department of Public Health Engineering (Bangladesh) undertaking a project funded by the UK Department for International Development (DFID). Any views expressed are not necessarily those of DFID'. In cases where only a map or diagram is reproduced or where data from the report are used, the above acknowledgement may be substituted by a full citation to the report as follows:

BGS and DPHE. 2001. Arsenic contamination of groundwater in Bangladesh. Kinniburgh, D G and Smedley, P L (Editors). British Geological Survey Technical Report WC/00/19. British Geological Survey: Keyworth.

Contents

List of Abbreviations

As	Arsenic
BGS	**B**ritish **G**eological **S**urvey
GP	**G**aussian **P**rocesses
GLM	**G**eneralized **L**inear **M**odel
MCMC	**M**arkov **c**hain **M**onte **C**arlo
M-H	**M**etroplis-**H**astings
HMC	**H**amiltonian **M**onte **C**arlo
NUTS	**N**o-**U**-**T**urn **S**ampler
WAIC	**W**atanabe-**A**kaike **I**nformation **C**riterion
WAIC	**W**idely **A**pplicable **I**nformation **C**riterion
LOO-CV	**L**eave-**o**ne-**o**ut **C**ross **V**aliation
UTM	**U**niversal **T**ransverse **M**ercator
RMSE	**R**oot **M**ean **S**quare **E**rror
EDA	**E**xploratory **D**ata **A**nalysis
GLGC	**G**aussian-**L**og **G**aussian **C**onvolution
NNGP	**N**earest **N**eighbor **G**aussian **P**rocesses

Chapter 1

Introduction

1.1 Background

Bangladesh is a country with a population of about 160 million (according to 2011 census). It is located in the Ganges, Brahmaputra and Meghna plains. A major part of the country is low lying (80%) and is being annually flooded. The country has a network of hundreds of rivers.

The arsenic contamination in drinkable water has caused serious problems to human health, it is estimated that 85 million people is at risk from arsenic (As) in drinking water and in food crops (Hossain, 2006). According to WHO [1], long-term exposure to arsenic from drinking-water and food can cause cancer and skin lesions. It has also been associated with cardiovascular disease and diabetes. In utero and early childhood exposure has been linked to negative impacts on cognitive development and increased deaths in young adults.

Comilla is one of the most polluted region (district) in Bangladesh. According to Flanagan, Johnston, and Zheng, 2012, Comilla has the highest number of arsenic-related deaths – 3748 adult deaths in 2009. Many people there are exposed to high arsenic concentrations. Resulting losses in productivity could be estimated up to US$ 1.1 billion over the next 20 years in Comilla alone.

Bangladesh's groundwater arsenic contamination and the related health effect came to notice in 1992 by researchers from Jadavpur University (India) and they officially informed the government of Bangladesh. The authorities of Bangladesh firstly identified the problem in 1993. Later in 1994, researchers in Jadavpur University also informed the situation to aid agencies (Chakraborti et al., 2015). A more comprehensive survey was carried out later in 1998.

The project "*Groundwater Studies for Arsenic Contamination in Bangladesh* " [2] began in January 1998 and has been funded throughout by the UK Department for International Development (DFID). The project was carried out in a collaboration between a

[1] https://www.who.int/news-room/fact-sheets/detail/arsenic
[2] Acknowledgment

number of organisations. On behalf of the Government of Bangladesh, DFID appointed the British Geological Survey (BGS) as lead consultants for the study. The Department of Public Health Engineering (DPHE), which is responsible for water supply throughout the country other than in the cities of Dhaka and Chittagong, was the executing agency. The Bangladesh Water Development Board (BWDB) and Geological Survey of Bangladesh (GSB) also provided counterparts.

The project was carried out in two phases: an initial six-month 'Rapid Investigation Phase' (Phase 1), and a subsequent 18-month Phase 2.

We examine the arsenic data from the district of Comilla gathered from the BGS survey and propose a statistical model to describe the distribution of arsenic levels across Comilla. Our preliminary analysis reveals that there are two major challenges for modeling the arsenic levels in Comilla. The arsenic level is highly right skewed and there exists left-censored (below detection limit) data in Comilla.

1.2 Literature Review

Environmental Perspective There was some preliminary reports released after the Phase 1 survey [3]. In Feburary 2001, BGS have published their Phase 2 report of the survey (Kinniburgh and Smedley, 2001), which has 4 volumes in total. All Phase 2 reports and data are publicly available.

Later, publications are also discussing these topics on the survey, such as Hossain, 2006, Flanagan, Johnston, and Zheng, 2012 and follow-up studies like Chakraborti et al., 2015. According to Chakraborti, the follow-up study shows that

1. Villagers are now more aware about the danger of drinking arsenic polluted water;
2. Villagers are currently drinking less arsenic contaminated water;
3. Many villagers in affected village died of cancer;
4. Arsenic contaminated water is in use for agricultural irrigation and arsenic exposure from food chain could be future danger.

Statistical Perspective There are a few works from the perspective of statistical models. Yu, Harvey, and Harvey, 2003 examined log-scale data of the whole country in an exploratory level by estimating an exponential spatial correlation structure through semi-variogram. Yu considering well depth and geological factor as covariates. However, for the censored data, Yu simply assigned a value of 0.1 $\mu g/L$.

[3] http://www.bgs.ac.uk/research/groundwater/health/arsenic/Bangladesh/reports.html

Gaus et al., 2003 used *Disjunctive kriging* to estimate concentrations of arsenic in the shallow ground water and to map the probability that the national limit for arsenic in drinking water was exceeded for most of the country.

Sen, 2016 proposed another way of measuring the pollution referred as the *Contamination Severity Index* (CSI), which is based on the Gini and Gastwirth coefficient. Through his proposal, observations are aggregated into areal data at different scales in sub-regions.

Skew-Data For skewed spatial data, Zhang and El-Shaarawi, 2010 proposed a skew process model in extension of the *skew-normal* distribution. Zhang and El Shaarawi estimated the parameters through the Monte Carlo EM algorithm.

Zareifard et al., 2018 proposed another skew process called *Gaussian-log Gaussian convolution* (GLGC) to construct latent spatial models which provide great flexibility in capturing skewness. We will have a short discussion on these two skew process models in Section 3.3.

Censoring Censored data have been widely studied. Militino and Ugarte, 1999 proposed an adaptation of the traditional methodology using the EM algorithm. The approach allows estimation when censoring is present.

Rathbun, 2006 applied the *Robbins-Monro algorithm* for estimating the parameters of a spatial regression model which uses importance sampling to obtain conditional samples of left-censored observations. A predictor for data at unsampled sites is obtained by taking the weighted mean of kriging predictors computed from independent importance samples.

1.3 Objective

The objective of this thesis is to propose a spatial model to describe the arsenic levels across the Comilla district. We will build up the models in the form of *Hierarchical Models*. The inference procedure will be performed under the Bayesian paradigm. Finally, we will use the model to predict the As levels for the whole region of Comilla.

Thesis is organized as follows.

Chapter 2 proposes the models and provides necessary information for performing inference, assessing performance between different models and making predictions for unobserved locations.

Chapter 3 runs a simulations study to determine some specification of the prior setting for the hierarchical models and use the specifications to model the Comilla's data and perform spatial interpretation (*kriging*). We also compare our model(s) with other spatial models and some recent models proposed in the literature.

Chapter 4 will review and conclude some finding through our study, bring up some limitations of the models and discuss some possible future work.

Chapter 2

Proposed Model

In our study, the data we examine is called *point-referenced data*, sometimes referred as geocoded or geostatistical data. It is one of three types of spatial data following Cressie, 1992. For point-referenced data, $Y(\mathbf{s})$ is random vector at location $\mathbf{s} \in \mathbb{R}^p$, for instance, $\mathbf{s} \in \mathbb{R}^2$. Here $\mathbf{s}$ varies continuously over $\mathcal{D}$, a fixed subset of $\mathbb{R}^p$ that contains an p-dimensional rectangle of positive volume (Banerjee, Carlin, and Gelfand, 2014). One of the most important goal in geostatistics is to make predictions on unobserved locations based on observed locations.

Point-referenced data can be thought of as resulting from observations on the stochastic process $\{Y(\mathbf{s}), \mathbf{s} \in \mathcal{D}\}$. We start from modeling the spatial process.

2.1 Model-Based Geostatistics

Gaussian Processes A stochastic process $S(\mathbf{s})$ is a *Gaussian process* (GP) if the joint distribution of any finite subset $S(\mathbf{s}_1), \ldots, S(\mathbf{s}_n)$ is a multivariate normal distribution for any integer n and locations $\mathbf{s}_i, i \in \{1, \ldots, n\}$. A GP is fully specified by the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, denoted as $\mathrm{GP}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Isotropy and Intrinsic Stationary One of the simplified assumption in geostatistics is called *isotropy*. The assumption assumes that the covariance or correlation of two locations $(\mathbf{s}_i, \mathbf{s}_j)$ only depends on their Euclidian distance $d = \|\mathbf{d}\| = \|\mathbf{s}_i - \mathbf{s}_j\|$. If it is not the case, we may say the process is anisotropic. We define the *covariance function* of two locations $(\mathbf{s}_i, \mathbf{s}_j)$ or $\{(\mathbf{s} + \mathbf{d}), \mathbf{s}\}$ as follows,

$$C(d) = C(\mathbf{d}) = \mathrm{cov}(S(\mathbf{s}_i), S(\mathbf{s}_j)) = \sigma^2 \rho(d) = \sigma^2 \rho \left\|\mathbf{s}_i - \mathbf{s}_j\right\| \tag{2.1}$$

A process $Y(\mathbf{s})$ is called *intrinsically stationary* if assuming

$$\mathbb{E}\left[Y(\mathbf{s} + \mathbf{d}) - Y(\mathbf{s})\right] = 0 \tag{2.2}$$

$$\text{define: } \mathbb{E}\left[Y(\mathbf{s} + \mathbf{d}) - Y(\mathbf{s})\right]^2 = \mathrm{var}\left[Y(\mathbf{s} + \mathbf{d}) - Y(\mathbf{s})\right] = 2\gamma(\mathbf{d}) \tag{2.3}$$

here the function $2\gamma(\mathbf{d})$ is called variogram and $\gamma(\mathbf{d})$ is called *semi-variogram*. Note that Equation (2.3) is valid only under the assumption of isotropic. It can be shown that variogram and covariance function has following relations,

$$\begin{aligned} 2\gamma(\mathbf{d}) &= \mathrm{var}(Y(\mathbf{s}+\mathbf{d}) - Y(\mathbf{s})) \\ &= \mathrm{var}\left[Y(\mathbf{s}+\mathbf{d})\right] + \mathrm{var}\left[Y(\mathbf{s})\right] - 2\mathrm{cov}\left[Y(\mathbf{s}+\mathbf{d}), Y(\mathbf{s})\right] \\ &= 2C(\mathbf{0}) - 2C(\mathbf{d}) \\ \Rightarrow \gamma(\mathbf{d}) &= C(\mathbf{0}) - C(\mathbf{d}) \end{aligned} \tag{2.4}$$

Under this assumption, Diggle, Tawn, and Moyeed, 1998 provides a general framework adapted the linear and generalized linear spatial models. For simplicity, in the following text, we will use $Z(\mathbf{s})$ to represent a zero mean GP with covariance matrix $\mathbf{\Sigma}$.

$$Y(\mathbf{s}) = m_{\boldsymbol{\theta}}(\mathbf{s}) + Z(\mathbf{s}) + \varepsilon(\mathbf{s}) \tag{2.5}$$

where $m_{\boldsymbol{\theta}}(\mathbf{s}) = \mathbf{x}(\mathbf{s})^\top \boldsymbol{\theta}$ is the mean effect, $Z(\mathbf{s})$ the latent Gaussian process capturing the spatial structure and $\varepsilon(\mathbf{s})$ are mutually independent $N(0, \tau^2)$. In geostatistics, τ^2 is often referred as *nugget effect*.

We can also describe Equation (2.5) as hierarchical model,

1. $\left\{Z(\mathbf{s}) : \mathbf{s} \in \mathbb{R}^2\right\}$ is a Gaussian process with mean $\mathbf{0}$ and covariance function $C(d)$.
2. Conditioning on $Z(\mathbf{s})$, the $Y(\mathbf{s})$ are independent realisations from a normal distribution with conditional means $m_{\boldsymbol{\theta}}(\mathbf{s}) = \mathbf{x}(\mathbf{s})^\top \boldsymbol{\theta}$ and variances τ^2.

To define a proper model, the correlation function $\rho(d)$ must be positive-definite between any pair of locations. In practice, we impose some parametric model for $\rho(d)$ to ensure the validity of the positive-definite condition. Maybe the simplest correlation function is known as *Exponential Correlation function*,

$$\rho(d) = \exp\left\{-d_{ij}/\phi\right\} \tag{2.6}$$

where $d = d_{ij} = \left\|\mathbf{s}_i - \mathbf{s}_j\right\|$ is the Euclidean distance between two locations and $\phi > 0$ is a scale parameter.

The idea of hierarchical model can be further extend into generalized linear models (GLM) by modifying the second condition and add one more for GLMs.

$$g\left\{\mathbb{E}\left[Y(\mathbf{s})\right]\right\} = \eta(\mathbf{s}) = m_{\boldsymbol{\theta}}(\mathbf{s}) + Z(\mathbf{s}) \tag{2.7}$$

the hierarchical formulation is as follows,

1. $\left\{Z(\mathbf{s}) : \mathbf{s} \in \mathbb{R}^2\right\}$ is a Gaussian process with mean $\mathbf{0}$ and covariance function $C(d)$.

2. Conditioning on $Z(\mathbf{s})$, the $Y(\mathbf{s}_i)$ are independent realisations from some distribution family with conditional means $m_{\boldsymbol{\theta}}(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)^\top \boldsymbol{\theta}$ and possibly some common parameter(s) common to all observations. The function $g\{\cdot\}$ depends on the distribution family.

3. $\eta(\cdot)$ are some link functions (e.g. canonical link).

We propose a spatial gamma model to the As levels in Comilla in next section.

2.1.1 Map Projection

As we discussed earlier, the correlation function is measured by the Euclidean distance. In practice, location information is recorded as longitude and latitude.

Here, we give a brief introduction to geometry of (and determine distances on) the surface of the earth before getting into any computations. The main purpose is to remind the reader about the difference of common used spherical representation (i.e. longitude and latitude) and Euclidean distance. Using spherical coordinates to represent distance may lead to distortion when locations are away from the equator.

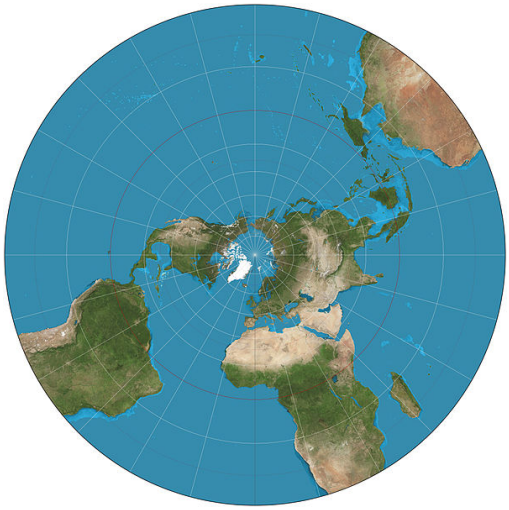
WGS84 The World Geodetic System (WGS) is a standard for use in cartography, geodesy, and satellite navigation including GPS. This standard includes the definition of the coordinate system's fundamental and derived constants, the ellipsoidal (normal) Earth Gravitational Model (EGM), a description of the associated World Magnetic Model (WMM), and a current list of local datum transformations. Some standard known parameters are

Name	Parameter	Value
Semi-major axis	a	6378137 m
Simi-minor axis	b	6356752 m
Flattening	f	1/298.257223563
Eccentricity	e	0.00669438

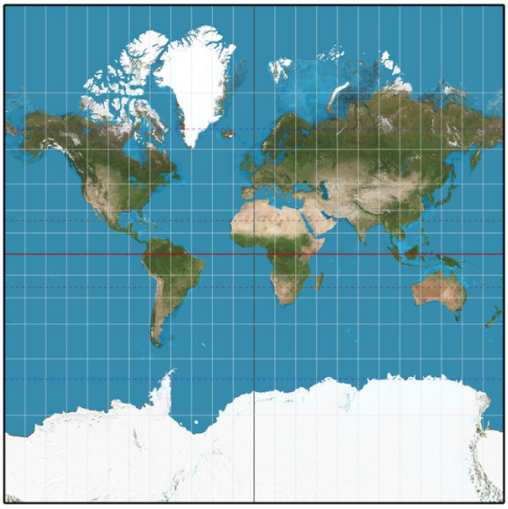
TABLE 2.1: WGS 84 Parameters

A map projection is a systematic representation of all or part of the surface of the earth on a plane. This typically comprises lines delineating meridians (longitudes) and parallels (latitudes), as required by some definitions of the projection. A well-known fact from topology is that it is impossible to prepare a distortion-free flat map of a surface curving in all directions (Banerjee, Carlin, and Gelfand, 2014).

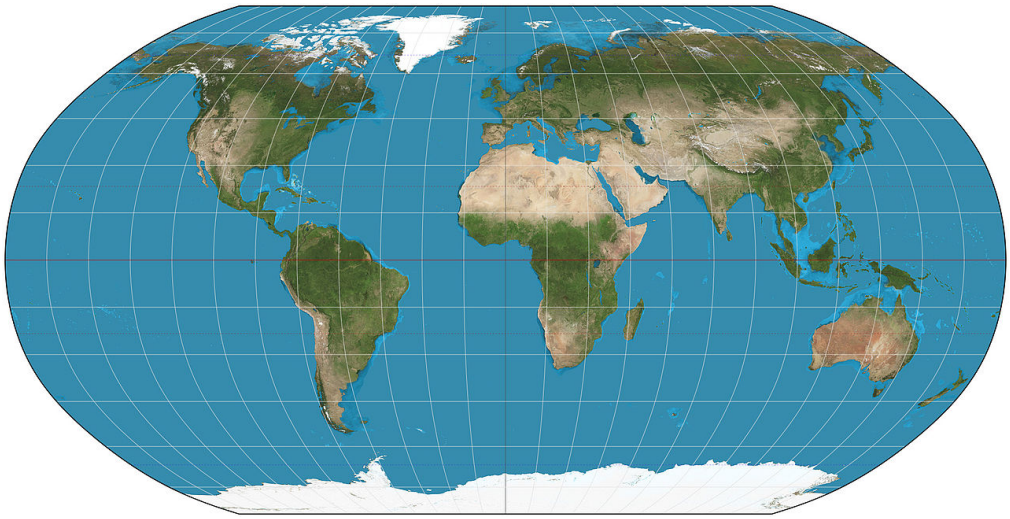
There are many map projections, each maintains certain characteristic(s) during projection, Figure 2.1 provides some example of projections.



(A) Stereographic



(B) Mercator



(C) Robinson

FIGURE 2.1: Commonly Used Map Projections, graphs taken from Wikipedia [Stereographic, Mercator, Robinson]. Stereographic is conformal, it preserves angles at which curves meet. Mercator is conformal and true direction. Robinson distort all attributes but create a "more pleasant" appearance

Universal Transverse Mercator The *Universal Transverse Mercator* (UTM) coordinate system is adopted by The National Imagery and Mapping Agency (NIMA) and used especially for military use throughout the world Banerjee, Carlin, and Gelfand, 2014.

UTM divides the Earth into 60 zones, each with 6°of longitude in width. Zone 1 covers longitude 180°to 174°W; zone numbering increases eastward to zone 60, which covers longitude 174°E to 180°. The polar regions south of 80°S and north of 84°N are excluded.

Each of the 60 zones uses a transverse Mercator projection that can map a region of large north-south extent with low distortion. By using narrow zones of 6°of longitude (up to 668 km) in width, and reducing the scale factor along the central meridian to $k_0 = 0.9996$ (a reduction of 1:2500), the amount of distortion is held below 1 part in 1,000 inside each zone. Distortion of scale increases to 1.0010 at the zone boundaries along the equator. In WGS84, the formula for converting (λ, ϕ) (in radian) to (x, y) (in meters) follows Snyder, 1987, page 61,

$$
\begin{aligned}
x &= k_0 N \left[A + \frac{(1 - T + C)A^3}{3!} + \frac{(5 - 18T + T^2 + 72C - 58e'^2)A^5}{5!} \right] \\
y &= k_0 \left\{ M - M_0 + N \tan\phi \left[\frac{A^2}{2} + \frac{(5 - T + 9C + 4C^2)A^4}{4!} + \frac{(61 - 58T + T^2 + 600C - 330e'^2)A^6}{6!} \right] \right\} \\
k &= k_0 \left[1 + \frac{(1 - C)A^2}{2!} + \frac{(5 - 4T + 42C + 13C^2 - 28e'^2)A^4}{4!} + \frac{(61 - 148T + 16T^2)A^6}{6!} \right]
\end{aligned}
$$

where

$$
\begin{aligned}
\phi_0 &= 0 \quad \text{latitude of the central meridian at the origin of the } x, y \text{ coordinates} \\
M_0 &= 0 \quad M \text{ at } \phi_0 \\
\lambda_0 &= \text{longitude of central meridian (for UTM zone)}
\end{aligned}
$$

and

$$
\begin{aligned}
k_0 &= 0.9996 \\
e'^2 &= e^2 / (1 - e^2) \\
N &= a / \sqrt{1 - e^2 \sin^2 \phi} \\
T &= \tan^2 \phi \\
C &= e'^2 \cos^2 \phi \\
A &= (\lambda - \lambda_0) \cos \phi
\end{aligned}
$$

and

$$M = a\left[\left(1 - \frac{e^2}{4} - \frac{3e^4}{64} - \frac{5e^6}{264} - \cdots\right)\phi - \left(\frac{3e^2}{8} + \frac{3e^4}{32} - \frac{45e^6}{1024} + \cdots\right)\sin 2\phi \right.$$
$$\left. + \left(\frac{15e^4}{8} + \frac{45e^6}{32} + \cdots\right)\sin 4\phi - \left(\frac{35e^6}{3072} + \cdots\right)\sin 6\phi + \cdots\right]$$

In practice, UTM is used by overlaying a transparent grid on the map, allowing distances to be measured in meters (we may divide distance under UTM by 1000 to measure it under kilometers) at the map scale between any map point and the nearest grid lines to the south and west.

For convenience, we may use some tools to help us easily identify the UTM Zone. For example, a map developed by Robertyoung from MangoMap or interactive map developed by ArcGIS. The transformation between geo-coordinates and UTM will also be handled by some software package, e.g. `sp` in `R`.

The rest of the text, we describe locations using *easting* and *northing* in kilometers.

2.2 Proposed Model

Gamma distribution is a two-parameter family of continuous probability distributions, it belongs to the exponential family. If we define $\alpha > 0$ as shape parameter and $\beta > 0$ as rate parameter, the density of gamma, denoted as $\text{Gamma}(\alpha, \beta)$, is

$$p(y|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\beta y), y > 0 \tag{2.8}$$

where $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-\alpha y} dy$ is the *Gamma function*.

The distribution function is

$$P(y|\alpha, \beta) = \int_0^y p(t|\alpha, \beta) dt = \frac{1}{\Gamma(\alpha)} \cdot \gamma(\alpha, \beta y) \tag{2.9}$$

where $\gamma(\cdot)$ is the *lower incomplete Gamma function*.

For this parameter setting,

$$\mathbb{E}(y) = \frac{\alpha}{\beta} \qquad \text{var}(y) = \frac{\alpha}{\beta^2}$$

Gamma Spatial Model Let $\mathbf{s}_i \in \mathbf{s} \subset \mathbb{R}^2$ be a specific location, we construct our spatial model through hierarchical models,

$$\begin{aligned} Y(\mathbf{s}_i) &\sim \text{Gamma}(\beta \cdot \alpha(\mathbf{s}_i), \beta) \\ \ln \alpha(\mathbf{s}_i) &= \mathbf{x}(\mathbf{s}_i)^\top \boldsymbol{\theta} + Z(\mathbf{s}_i) \\ Z(\mathbf{s}) &\sim \text{GP}(\mathbf{0}, \boldsymbol{\Sigma}) \end{aligned} \tag{2.10}$$

where we use the exponential correlation structure, the variance-covariance matrix is $\Sigma_{ij} = \sigma^2 \exp(-d_{ij}/\phi)$. Figure 2.2 also illustrates Equation (2.10),

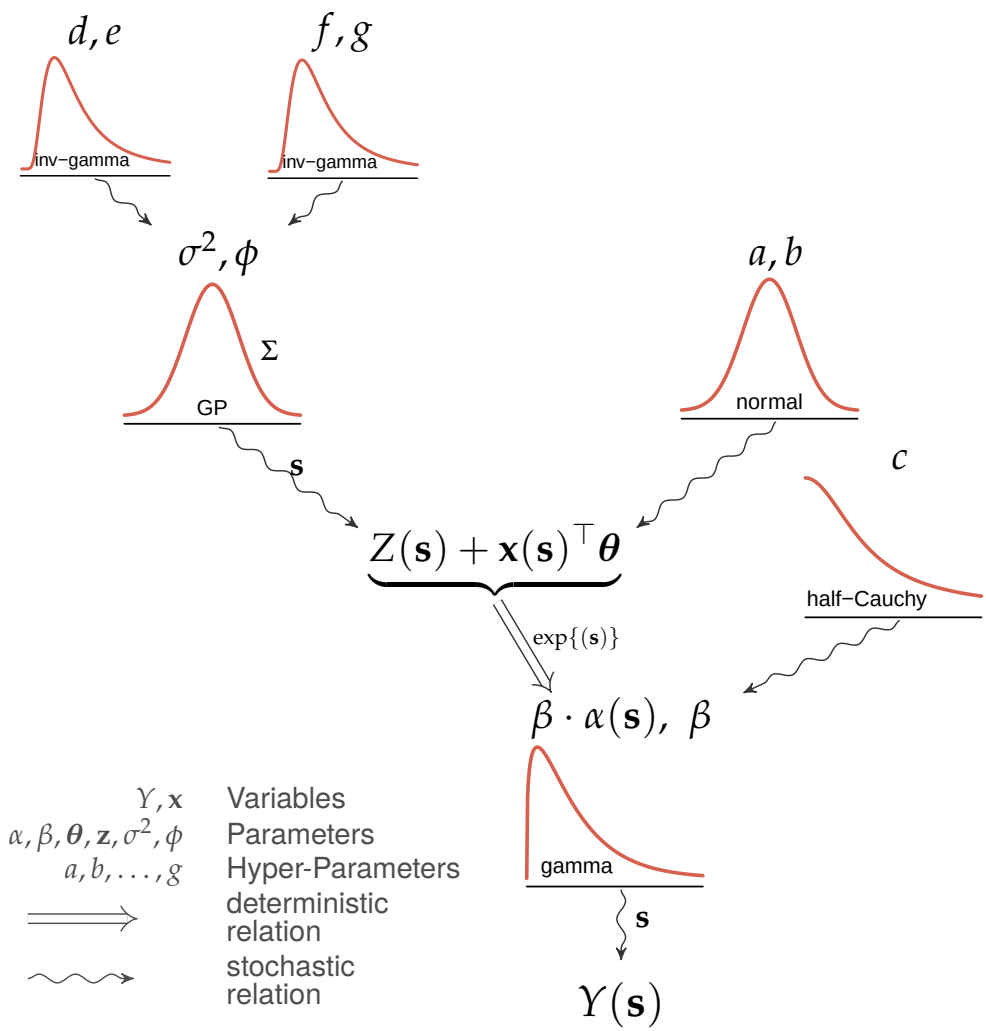


FIGURE 2.2: Kruschke Style Diagrams of Equation (2.10). Example inspired by Kruschke, 2014. Distribution plot credit to Rasmus Bååth; LaTeX TikZ example credit to Tinu Schneider

2.3 Properties

For Equation (2.10), to investigate the marginal mean, variance and covariance of $Y(\mathbf{s})$, if we integrate $Y(\cdot)$ with respect to $Z(\cdot)$, we may not obtained a closed form solution. However, we know that $Y(\mathbf{s})|Z(\mathbf{s})$ follows gamma distribution, and $Z(\mathbf{s})$ follows log-normal distribution. Applying the law of total expectation, variance, covariance and using the properties of log-normal, we have the marginal properties of the gamma spatial model.

$$\begin{aligned}
\mathbb{E}\left\{Y(\mathbf{s}_i)\right\} =& \mathbb{E}\left\{\mathbb{E}\left[Y(\mathbf{s}_i)|Z(\mathbf{s}_i)\right]\right\} \\
=& \mathbb{E}\left\{\alpha(\mathbf{s}_i)\right\} \\
=& \mathbb{E}\left\{\exp\left[\mathbf{x}(\mathbf{s}_i)^\top\boldsymbol{\theta} + Z(\mathbf{s}_i)\right]\right\} \\
=& \exp\left\{\mathbf{x}(\mathbf{s}_i)^\top\boldsymbol{\theta}\right\} \cdot \mathbb{E}\left\{\exp\left[Z(\mathbf{s}_i)\right]\right\} \\
=& \exp\left\{\mathbf{x}(\mathbf{s}_i)^\top\boldsymbol{\theta} + \sigma^2/2\right\}
\end{aligned} \tag{2.11}$$

$$\begin{aligned}
\mathrm{var}\left\{Y(\mathbf{s}_i)\right\} =& \mathbb{E}\left\{\mathrm{var}\left[Y(\mathbf{s}_i)|Z(\mathbf{s}_i)\right]\right\} + \mathrm{var}\left\{\mathbb{E}\left[Y(\mathbf{s}_i)|Z(\mathbf{s}_i)\right]\right\} \\
=& \mathbb{E}\left\{\alpha(\mathbf{s}_i)/\beta\right\} + \mathrm{var}\left\{\alpha(\mathbf{s}_i)\right\} \\
=& \beta^{-1}\exp\left\{\mathbf{x}(\mathbf{s}_i)^\top\boldsymbol{\theta} + \sigma^2/2\right\} + \exp\left\{2\mathbf{x}(\mathbf{s}_i)^\top\boldsymbol{\theta} + \sigma^2/2\right\}\left[\exp(\sigma^2) - 1\right]
\end{aligned} \tag{2.12}$$

$$\begin{aligned}
\mathrm{cov}\left\{Y(\mathbf{s}_i), Y(\mathbf{s}_j)\right\} =& \mathbb{E}\left\{\mathrm{cov}[Y(\mathbf{s}_i), Y(\mathbf{s}_j)|Z(\mathbf{s})]\right\} + \mathrm{cov}\left\{\mathbb{E}[Y(\mathbf{s}_i)|Z(\mathbf{s}_i)], \mathbb{E}[Y(\mathbf{s}_j)|Z(\mathbf{s}_j)]\right\} \\
=& \mathrm{cov}\left\{\alpha(\mathbf{s}_i), \alpha(\mathbf{s}_j)\right\} + 0 \\
=& \mathbb{E}(\alpha(\mathbf{s}_i))\mathbb{E}(\alpha(\mathbf{s}_j))\left[\exp(\rho\sigma_i\sigma_j) - 1\right] \\
=& \exp\left\{\mathbf{x}(\mathbf{s}_i)^\top\boldsymbol{\theta} + \mathbf{x}(\mathbf{s}_j)^\top\boldsymbol{\theta} + \sigma^2\right\}\left\{\exp\left[\sigma^2 \cdot \rho(d_{ij})\right] - 1\right\}
\end{aligned} \tag{2.13}$$

where $\rho(d_{ij})$ is the correlation function of $Z(\mathbf{s})$.

2.4 Inference Procedure

In this study, the inference will be performed under the Bayesian paradigm. Let $\mathbf{y} = (\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{cen}})^\top$ be the vector of observations, where $\mathbf{y}_{\text{obs}}$ is the measured values and $\mathbf{y}_{\text{cen}}$ is the left censored data, we only know that the As is below certain levels L_{cen}. Using the conditional independence of $Y(\mathbf{s})|Z(\mathbf{s})$ the likelihood is given by

$$
\begin{aligned}
p(\mathbf{y}|\alpha(\mathbf{s}),\beta) = &\left\{\prod_{i=1}^{n_{\text{obs}}} \frac{\beta^{\beta\cdot\alpha(\mathbf{s}_i)}}{\Gamma(\beta\cdot\alpha(\mathbf{s}_i))} y_i^{\beta\cdot\alpha(\mathbf{s}_i)-1}\exp(-\beta y_i)\right\} \times \\
&\left\{\prod_{i=1}^{n_{\text{cen}}} \frac{1}{\Gamma(\beta\cdot\alpha(\mathbf{s}_i))}\cdot\gamma(\beta\cdot\alpha(\mathbf{s}_i),\beta L_{\text{cen}})\right\}
\end{aligned}
\tag{2.14}
$$

The second part of Equation (2.14), i.e. likelihood for $\mathbf{y}_{\text{cen}}$ is assigned as distribution functions (Gelman, 2004). If we have more than one types of censoring, we may further split $\mathbf{y}_{\text{cen}}$ into product of two or more different distribution functions, for instance, for two types of censoring, we assign

$$
P(\mathbf{y}_{\text{cen}}|\alpha(\mathbf{s}),\beta) = \prod_{i=1}^{n_{\text{cen1}}} \frac{\gamma(\beta\cdot\alpha(\mathbf{s}_i),\beta L_{\text{cen1}})}{\Gamma(\beta\cdot\alpha(\mathbf{s}_i))}\cdot\prod_{i=1}^{n_{\text{cen2}}} \frac{\gamma(\beta\cdot\alpha(\mathbf{s}_i),\beta L_{\text{cen2}})}{\Gamma(\beta\cdot\alpha(\mathbf{s}_i))}
\tag{2.15}
$$

The model specification will be complete once all the prior distributions for the parameters and hyper-parameters are assigned. We assign a single prior distribution to the parameter vector. For convenience, we assume independence among some parameters, thus for some parameters the prior can be expressed as product of the prior distributions.

The log scale of spatial structure is a GP, the prior distribution of $Z(\mathbf{s})$ is

$$
p(Z(\mathbf{s})|\boldsymbol{\mu}=\mathbf{0},\boldsymbol{\Sigma}) = (2\pi)^{-n/2}\det(\boldsymbol{\Sigma})^{-1/2}\exp\left[-\frac{1}{2}\mathbf{z}^\top\boldsymbol{\Sigma}^{-1}\mathbf{z}\right]
\tag{2.16}
$$

where $n = n_{\text{obs}} + n_{\text{cen}}$ and $Z(\mathbf{s}) = \mathbf{z} = (\mathbf{z}_{\text{obs}}, \mathbf{z}_{\text{cen}})^\top = (z_1,\ldots,z_n)^\top$.

For $\{\boldsymbol{\theta},\beta,\sigma^2,\phi\}$, we assign independent priors to each of the parameters. Combining Equation (2.14) and (2.16), and the priors of $\{\boldsymbol{\theta},\beta,\sigma^2,\phi\}$, the posterior distribution to be sampled from is given by

$$
\begin{aligned}
p(\boldsymbol{\theta},\beta,\mathbf{z},\sigma^2,\phi|\mathbf{y}) \propto &\left\{\prod_{i=1}^{n_{\text{obs}}} \frac{\beta^{\beta\cdot\alpha(\mathbf{s}_i)}}{\Gamma(\beta\cdot\alpha(\mathbf{s}_i))} y_i^{\beta\cdot\alpha(\mathbf{s}_i)-1}\exp(-\beta y_i)\right\} \times \\
&\left\{\prod_{i=1}^{n_{\text{cen}}} \frac{1}{\Gamma(\beta\cdot\alpha(\mathbf{s}_i))}\cdot\gamma(\beta\cdot\alpha(\mathbf{s}_i),\beta L_{\text{cen}})\right\} \times \\
&\det(\boldsymbol{\Sigma})^{-1/2}\exp\left[-\frac{1}{2}\mathbf{z}^\top\boldsymbol{\Sigma}^{-1}\mathbf{z}\right] \times \\
&\prod_{i=0}^{k} p(\theta_i) \times p(\beta)p(\sigma^2)p(\phi)
\end{aligned}
\tag{2.17}
$$

where $\boldsymbol{\theta} = (\theta_0,\theta_1,\ldots,\theta_k)^\top$ corresponds to one intercept and k covariates. $p(\boldsymbol{\theta},\beta,\mathbf{z},\sigma^2,\phi|\mathbf{y})$ is called the *kernel* of the posterior distribution, it is proportional to the density of the

posterior distribution.

One of the advantages of using Bayesian paradigm is that we can build such complex hierarchical models by simply stacking on likelihoods functions and corresponding prior distributions for the hyper-parameters, the uncertainty about the parameters will be automatically accounted for.

The posterior distribution (Equation 2.17) is complex, and a closed form solution may not be able to obtained, hence under Bayesian framework, we will obtain the parameter vector using *Markov chain Monte Carlo* (MCMC) methods.

Markov Chain Monte Carlo Soon after Monte Carlo method were invented, MCMC were proposed by Metropolis et al., 1953. The principle is to construct a Markov chain for the posterior distribution $\boldsymbol{\pi}$ such that the chain has stationary distribution (i.e. independent with respect to initial state after time t) and the transition probabilities of the chain have simple form. Finally, the chain has to be ergodic to ensure that every states is able to be visited.

Two algorithms for MCMC are quite popular. The *Metropolis-Hastings* (M-H) algorithm is an adaptation of a random walk with an acceptance/rejection rule to converge to the specified target distribution. It starts from some initial state in the space, at each step the acceptance/rejection decision is applied by determine probability of jumping via comparing the density of the proposal and target distribution. Suppose we want to generate sample vectors $\boldsymbol{\theta}$ of k dimensions. The algorithm of the M-H with proposal distribution $q(\cdot|\cdot)$ can be described in Algorithm 1,

Algorithm 1: Metropolis Hastings Algorithm

Input: Proposal distribution
Output: Samples from target distribution

1 Initialize the chain at random place $\boldsymbol{\theta}^{(0)}$
2 **for** $\ell = 1$ **to** L **do**
3 Generate $\boldsymbol{\theta}^{\star}$ from proposal distribution $q(\boldsymbol{\theta}^{\star}|\boldsymbol{\theta}^{(\ell-1)})$
4 Compute the ratio

$$r = \frac{p(\boldsymbol{\theta}^{\star}|\mathbf{y})}{p(\boldsymbol{\theta}^{(\ell-1)}|\mathbf{y})} \cdot \frac{q(\boldsymbol{\theta}^{(\ell-1)}|\boldsymbol{\theta}^{\star})}{q(\boldsymbol{\theta}^{\star}|\boldsymbol{\theta}^{(\ell-1)})}$$

5 Set

$$\boldsymbol{\theta}^{(\ell)} = \begin{cases} \boldsymbol{\theta}^{\star} & \text{with probability } \min(r, 1) \\ \boldsymbol{\theta}^{(\ell-1)} & \text{otherwise} \end{cases}$$

The M-H algorithm degenerates to the *Metropolis* algorithm if the proposal distribution is symmetric, i.e. $q(\theta_a|\theta_b) = q(\theta_b|\theta_a)$, and for Metropolis algorithm, the ratio in line 4 of Algorithm 1 reduces to

$$r = \frac{p(\boldsymbol{\theta}^{\star}|\mathbf{y})}{p(\boldsymbol{\theta}^{(v-1)}|\mathbf{y})}$$

The performance of M-H algorithm depends on the proposal distribution $q(\cdot|\cdot)$. If the spread of $q(\cdot|\cdot)$ is too small, it will stuck in a very small area for a fairly long time. However, if the variation of $q(\cdot|\cdot)$ too large, it may lead to a fairly low acceptance rate while traversing.

The *Gibbs Sampler* is a special case of the M-H. Instead of moving in a high dimensional space in one shot, Gibbs sampler uses the posterior full conditional to break one movement into several movements in lower dimensions. Suppose we can divide a k-dimensional vector into m components, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^{\top}, m \leq k$. For a single jump in k-dimensional space, we cycle through the sub-vectors of $\boldsymbol{\theta}$, drawing each subset conditioning on the value of all the others, the jump is completed if all m components are sampled, in such case, the acceptance ratio for each component is 1. Algorithm 2 describes how Gibbs sampler works,

Algorithm 2: Gibbs Sampler Algorithm

Input: Posterior full conditional distributions
Output: Samples from target distribution

Initialize the chain at random place $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_m^{(0)})$
for $\ell = 1$ **to** L **do**
  Sample $\theta_1^{(\ell)}$ from $\mathbb{P}\left\{\theta_1 | \theta_2^{(\ell-1)}, \ldots, \theta_m^{(\ell-1)}\right\}$
  Sample $\theta_2^{(\ell)}$ from $\mathbb{P}\left\{\theta_2 | \theta_1^{(\ell)}, \ldots, \theta_m^{(\ell-1)}\right\}$
  Sample $\theta_i^{(\ell)}$ from $\mathbb{P}\left\{\theta_i | \theta_1^{(\ell)}, \theta_{i-1}^{(\ell)}, \theta_{i+1}^{(\ell-1)} \ldots, \theta_m^{(\ell-1)}\right\}, i = 1, 2, \ldots, m$

The basic M-H and Gibbs sampler algorithm can be seen as building blocks for more advanced Markov chain simulations. The Gibbs sampler is efficient when parameterized in terms of independent components. A highly dependent components that create slow convergence.

One possible solution is to apply transformations and reparameterization. For instance, performing a linear transformation of the parameters, but posterior distributions that are not approximately normal may require special methods. One of the method is called *Hamiltonian Monte Carlo* (HMC).

Hamiltonian Monte Carlo HMC borrows an idea from physics to suppress the local random walk behavior in the Metropolis algorithm, thus allowing it to move much more rapidly through the target distribution, it can also be viewed as another special case of the M-H algorithm. In HMC, for each component of $\theta_j \in \boldsymbol{\theta}$ we add an auxiliary variable

$\psi_j \in \boldsymbol{\psi}$ of same dimension. Both $\boldsymbol{\theta}, \boldsymbol{\psi}$ will be updated together during Metropolis steps, the jumping distribution for $\boldsymbol{\theta}$ is dominated by $\boldsymbol{\psi}$.

In physics, we describe the state of a particle by its position $\boldsymbol{\theta}$ and momentum $\boldsymbol{\psi}$. In Hamiltonian system, the total energy of a particle $\mathcal{H}(\boldsymbol{\theta}, \boldsymbol{\psi})$ is determined by kinetic and potential energy which can be written as

$$\mathcal{H}(\boldsymbol{\theta}, \boldsymbol{\psi}) = U(\boldsymbol{\theta}) + K(\boldsymbol{\psi}) \tag{2.18}$$

In statistics, an analogy can be drawn in Table 2.2, where $p_N(\mathbf{0}, \mathbf{M})$ is density of multivariate normal and $\mathbf{M}$ is pre-specified (known). The analogy is appropriate between the potential energy in and negative log density of the posterior via the concept of a *canonical distribution* from statistical mechanics (Brooks et al., 2011, Section 5.3.1).

In statistical mechanics, given some energy function $\mathcal{E}(x)$ for the state x of some physical system, the canonical distribution over states has probability density function

$$p(x) \propto \exp\{-\mathcal{E}(x)/\mathcal{T}\} \tag{2.19}$$

where $\mathcal{T}$ is the temperature of the system. View this in the opposite way, if we are interested in some distribution with density $p(x)$, we can obtain it as a canonical distribution with $\mathcal{T} = 1$ by setting $\mathcal{E}(x) = -\ln p(x)$. We set the energy function $\mathcal{E}(x)$ as Equation (2.18), Equation (2.19) becomes

$$\begin{aligned} p(\boldsymbol{\psi}, \boldsymbol{\theta}) &\propto \exp\{-\mathcal{H}(\boldsymbol{\theta}, \boldsymbol{\psi})\} \\ &\propto \exp\{-U(\boldsymbol{\theta})\} \exp\{-K(\boldsymbol{\psi})\} \end{aligned}$$

Since it can be written as product, this implies that $\boldsymbol{\theta}, \boldsymbol{\psi}$ are independent. Moreover, each of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ have the canonical distributions with energy function $U(\boldsymbol{\theta})$ and $K(\boldsymbol{\psi})$ respectively.

	Physics	Statistics
$\boldsymbol{\theta}$	position	values of parameters
$\boldsymbol{\phi}$	momentum	auxiliary variables
$U(\boldsymbol{\theta})$	potential energy	$-\ln(p(\boldsymbol{\theta}\|\mathbf{y}))$
$K(\boldsymbol{\psi})$	kinetic energy	$p_N(\mathbf{0}, \mathbf{M})$

TABLE 2.2: Analogy of physics and statistics for Hamiltonian system

Moreover, the three properties of Hamiltonian dynamics can also draw an analogy to Markov chain, see Table 2.3.

Because of the conservation of energy in Hamiltonian system, in a movement of particles, the increment of kinetic energy will result in the decrement of potential energy

Physics	Statistics
reversibility	equilibrium
conservation of energy	acceptance rate = 1
volume preservation symplecticness	volume of acceptance rate invariant

TABLE 2.3: Analogy of properties in Hamiltonian dynamics

and vice versa so that $\mathcal{H}(\boldsymbol{\theta}, \boldsymbol{\psi})$ will be unchanged. Using the second property in Table 2.3, an unchanged energy implies the acceptance rate equals 1. The change of $(\boldsymbol{\theta}, \boldsymbol{\psi})$ is described by a system of partial differential equations.

In practice, we measure the change of $(\boldsymbol{\theta}, \boldsymbol{\psi}) \Rightarrow \{U(\boldsymbol{\theta}), K(\boldsymbol{\psi})\}$ through numerical approximation for a series of small jumps of length ε. The traditional approximation includes Euler's method and Modified Euler's method (Brooks et al., 2011, Section 5.2.3), these two methods measure $\psi_i(\ell + \varepsilon)$ and $\theta_i(\ell + \varepsilon)$ directly. A better solution is obtained by a technique called *The Leapfrog Method*, the leapfrog breaks the measurement of ψ_i into even smaller piece,

$$\begin{aligned}
\psi_i(\ell + \varepsilon/2) &= \phi_i(\ell) - (\varepsilon/2) \cdot \frac{\partial U}{\partial \theta_i}(\boldsymbol{\theta}(\ell)) \\
\theta_i(\ell + \varepsilon) &= \theta_i(\ell) + \varepsilon \cdot \psi_i(\ell + \varepsilon/2) \\
\psi_i(\ell + \varepsilon) &= \psi_i(\ell + \varepsilon/2) - (\varepsilon/2) \cdot \frac{\partial U}{\partial \theta_i}(\boldsymbol{\theta}(\ell + \varepsilon))
\end{aligned} \tag{2.20}$$

Using the analogy from Table 2.2, the leapfrog method (Equation 2.20) can also be expressed as (Gelman et al., 2013)

$$\begin{aligned}
\boldsymbol{\psi} &\leftarrow \boldsymbol{\psi} + \frac{\varepsilon}{2} \cdot \frac{d \ln(p(\boldsymbol{\theta}|\mathbf{y}))}{d\boldsymbol{\theta}} \\
\boldsymbol{\theta} &\leftarrow \boldsymbol{\theta} + \varepsilon \mathbf{M}^{-1} \boldsymbol{\psi} \\
\boldsymbol{\psi} &\leftarrow \boldsymbol{\psi} + \frac{\varepsilon}{2} \cdot \frac{d \ln(p(\boldsymbol{\theta}|\mathbf{y}))}{d\boldsymbol{\theta}}
\end{aligned} \tag{2.21}$$

In this case, the total energy $\mathcal{H}(\boldsymbol{\theta}, \boldsymbol{\phi})$ may change a little due to the approximation of leapfrog method, going back to the analogy of statistics in Table 2.3, we know that HMC using leapfrog method ensures a high acceptance rate through a fairly long jump.

When moving through parameter space, we evaluate $U(\boldsymbol{\theta}), K(\boldsymbol{\psi})$ in a given length of time, it is equivalent of evaluating after fixed number of ε steps, then measure the change of total energy to determine the transition probability and lastly decide whether accept or reject the move.

Algorithm 3 describes the steps for obtaining random samples through HMC,

Algorithm 3: Hamiltonian Monte Carlo Algorithm

Input: covariance matrix $\mathbf{M}$
Output: Samples from target distribution

1 Initialize the chain at random place $\boldsymbol{\theta}^{(0)}$
2 **for** $\ell = 1$ **to** L **do**
3 Assign a random momentum $\psi \sim N(\mathbf{0}, \mathbf{M})$
4 Simulate the movement across energy surface (negative log probability) for a fixed time T (equivalent to a fixed steps $\mathcal{L}$)
5 Compute the ratio

$$r = \frac{p(\boldsymbol{\theta}^{\star}|\mathbf{y})}{p(\boldsymbol{\theta}^{(\ell-1)}|\mathbf{y})} \cdot \frac{p(\boldsymbol{\psi}^{\star})}{p(\boldsymbol{\psi}^{(\ell-1)})}$$

where $(\boldsymbol{\theta}^{(\ell-1)}, \boldsymbol{\psi}^{(\ell-1)})$ denotes the parameters at time 0 within each iteration, $(\boldsymbol{\theta}^{\star}, \boldsymbol{\psi}^{\star})$ denotes the parameters at time T (after $\mathcal{L}$ steps)
6 Set

$$\boldsymbol{\theta}^{(\ell)} = \begin{cases} \boldsymbol{\theta}^{\star} & \text{with probability } \min(r, 1) \\ \boldsymbol{\theta}^{(\ell-1)} & \text{otherwise} \end{cases}$$

Comparing HMC with M-H and Gibbs, the auxiliary variables allow HMC need neither to stay in a small jump to maintain a high acceptance rate Markov chain which restrict the performance of M-H nor avoiding inefficient when the components of the parameter vector are highly correlated in the case of Gibbs.

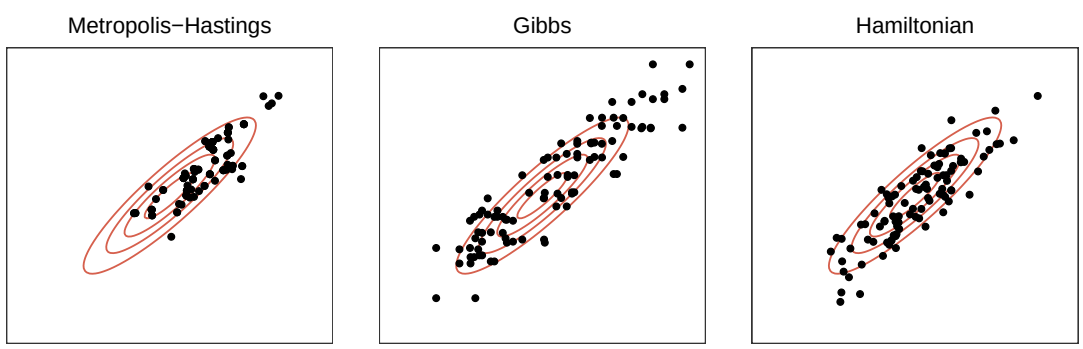


FIGURE 2.3: Comparison of MCMC Algorithm on bi-variate normal for first 100 sample draws, contours represent the density of the distribution. Example inspired by Hoffman and Gelman, 2014

Figure 2.3 illustrates a comparison of sampling a bi-variate normal through different MCMC algorithms. The bi-variate normal $\mathbf{x}$ with

$$\mathbf{x} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right)$$

We can see that for M-H algorithm, 100 samples does not spread over the distribution contour, it seems stuck in the right hand side of the contour. Gibbs sampler has a better performance already, but we may notice that sample spread more on the 'tail' than the middle. For HMC, the same number of samples spread quite good across the whole distribution contour.

It is worth mentioning that for the original HMC algorithm (Algorithm 3), the length of the time interval $[0, T]$ or the number of leapfrog steps $\mathcal{L}$ and the step size ε need to be carefully chosen. Sometimes, tuning $(\varepsilon, \mathcal{L})$ is more difficult than choosing proposal distributions in M-H (Brooks et al., 2011, Section 5.4.2). Hoffman and Gelman, 2014 proposed the *No-U-Turn sampler* (NUTS) algorithm. NUTS can determine the optimal time interval (or equivalently leapfrog steps $\mathcal{L}$). NUTS further improves the performance of HMC. For further detail of MCMC and HMC, please refer to Brooks et al., 2011; Betancourt, 2017.

Diagnostics As we stated in the beginning of MCMC, we construct the Markov chain such that the posterior distribution is stationary after time t. The samples obtained before time t is known as *warm-up* (burn-in) samples, they will be discarded since they are not good representatives of the stationary distributions. We use the post warm-up samples to describe distributions of the parameters. In practice, we do not know the value of t, it depends on the kernel of the posterior, the algorithm we use to obtain samples and the initial state of the parameters. Therefore, it is necessary to use some diagnostics to check that if the Markov chain is stationary [1]. In other word, if different initial states yield a similar distribution after some time t, we may say that the different Markov chains converge. There is no universal way to check if Markov chains have already converged, but there are many ways to provide evidences to show if the chains are likely to converge. We introduce two useful tools: *trace-plot* and *potential scale reduction factor on split chains* where we split each of the post warm-up chain in half and check the related results.

Trace-plots involve simulating two or more chains of parameter values and plotting the values of each chain against the sample number of the sampling process, typically on the same set of axes. If the chains are all representative of the posterior distribution, they should overlap each other and be unrelated to their randomly set starting positions. Moreover, the distributions for post warm-up samples of different Markov chains will roughly overlap since they are all good representatives of the parameters.

The potential scale reduction factor on split chains, denoted as $\hat{R}$ is proposed by Gelman and Rubin, 1992 and updated by Gelman et al., 2013, Section 11.4. The $\hat{R}$ statistic measures the ratio of the average variance of samples within each chain to the variance of the pooled samples across chains. If all chains are at equilibrium, these will be the

[1] There is another challenge in diagnostics known as label-switching, we will omit this in this thesis.

same and $\hat{R}$ will be one. If the chains have not converged to a common distribution, the $\hat{R}$ statistic will be greater than one.

Suppose for a scalar estimand ξ, if we label the simulations as ξ_{ij}, where $i = 1, \ldots, n$ is the number of iterations and $j = 1, \ldots, m$ is the number of Markov chains, define

$$\bar{\xi}_{.j} = \frac{1}{n}\sum_{i}^{n}\xi_{ij} \qquad \bar{\xi}_{..} = \frac{1}{m}\sum_{j}^{m}\bar{\xi}_{.j}$$

then define B, between-sequence variance and W, within-sequence variances.

$$B = \frac{n}{m-1}\sum_{j}^{m}(\bar{\xi}_{.j} - \bar{\xi}_{..})^2 \qquad W = \frac{1}{m(n-1)}\sum_{j}^{m}\sum_{i}^{n}(\xi_{ij} - \bar{\xi}_{.j})^2$$

then define Equation (2.22) follows the updated version by Gelman et al., 2013

$$\hat{\text{var}}^{+}(\xi|\mathbf{y}) = \frac{n-1}{n}W + \frac{1}{n}B \tag{2.22}$$

Finally, we have

$$\hat{R} = \sqrt{\frac{\hat{\text{var}}^{+}(\xi|\mathbf{y})}{W}} \tag{2.23}$$

this estimate declines to 1 as $n \to \infty$. If the potential scale reduction is high, then we have reason to believe that we may need more iteration to reach stationarity.

2.4.1 Implementation

As we have discussed, once we have kernel of the posterior distribution, we may be able to obtain samples of posterior parameters through MCMC. We may code the MCMC method by ourselves. Alternatively, in this study, we will use `Stan` platform to perform MCMC.

Stan Sampling through adaptive neighborhoods (Gelman et al., 2013, Section 12.6). `Stan`[2] is a state-of-the-art platform for statistical modeling and high-performance statistical computation. Users specify log density functions in `Stan`'s probabilistic programming language and get:

- Full Bayesian statistical inference with MCMC sampling (NUTS, HMC)
- Approximate Bayesian inference with variational inference (ADVI)
- Penalized maximum likelihood estimation with optimization (L-BFGS)

[2] https://mc-stan.org/

`Stan`'s math library provides differentiable probability functions & linear algebra (`C++ autodiff`). Additional `R` packages provide expression-based linear modeling, posterior visualization (`bayesplot`) and leave-one-out cross-validation (`loo`).

We will use `rstan` (version 2.19.2), the `R` (version 3.6.1) interface of `Stan` for this study. The compilation tools is `Rtools`[3] (version 3.5).

2.4.2 Prior Specifications

For $\boldsymbol{\theta}$, we can assign relatively vague priors, say a normal prior with fairly large variance, e.g. $\theta_i \sim N(0, 10^2), i = 0, \ldots, k$.

For ϕ, the parameter in the exponential correlation function, we consider assigning a Inverse Gamma distribution $\text{InvGamma}(a, b)$, for the value of the prior (a, b), we follow the suggestion of Schmidt, Gonçalves, and Velozo, 2017 and Banerjee, Carlin, and Gelfand, 2014, Section 2.1.3. For a, in inverse gamma the infinite variance implies that $a = 2$, for b using the notion *effective range*, where the correlation is negligible (say, 0.05) at half of the maximum distance $(d_{\max}/2)$. Apply this idea for the exponential correlation function,

$$0.05 = \exp\left(-\frac{d_{\max}}{2b}\right) \Rightarrow b \approx \frac{d_{\max}}{6} \tag{2.24}$$

hence we express the choice of prior for $\phi \sim \text{InvGamma}(2, d_{\max}/6)$.

From Equation (2.11) and (2.12), we know that the marginal mean and variance of the gamma model are contributed by the scale parameters β and σ, we have to carefully choose these priors to see if the posteriors distributions are sensitive to the prior specifications.

Gelman, 2006 suggests that instead of using non-informative prior distributions from the Inverse Gamma, e.g. $\text{InvGamma}(\varepsilon, \varepsilon)$, where $\varepsilon \to 0$ can take small values like 0.001, it may be more appropriate to use a weakly informative prior such as Uniform or half-Cauchy. Polson and Scott, 2012 also argues that half-Cauchy should be used as default prior for a top-level scale parameter in hierarchical models.

2.5 Model Comparison Criteria

Gelman, Hwang, and Vehtari, 2014 reviews and discusses various evaluation metric for comparing Bayesian models. In general, *Watanabe-Akaike Information Criterion* (WAIC) and Leave-one-out Cross-Validation (LOO-CV) are preferable among researchers. Gelman stated that WAIC and LOO-CV are asymptotically equivalent.

[3] `https://cran.r-project.org/bin/windows/Rtools/`

These two methods both use point-wise log predictive density to estimate out-of-sample prediction accuracy, a compromise from the expected log point-wise predictive density (Gelman et al., 2013, Section 7.1). For $\mathbf{y} = (y_1, \ldots, y_n)^\top$,

$$\begin{aligned}
\text{lppd} &= \text{log point-wise predictive density} \\
&= \sum_{i=i}^{n} \ln p_{\text{posterior}}(y_i) \\
&= \sum_{i=i}^{n} \ln \int p(y_i|\boldsymbol{\theta}) p(\theta|\mathbf{y}) d\boldsymbol{\theta} \\
&\approx \sum_{i=1}^{n} \ln \left\{ \frac{1}{L} \sum_{\ell=1}^{L} p(y_i|\boldsymbol{\theta}^{(\ell)}) \right\}
\end{aligned} \tag{2.25}$$

The last line is an approximation through simulation for $\ell = 1, \ldots, L$.

WAIC WAIC is proposed by Watanabe, 2010, also known as *Widely Applicable Information Criterion*. It is based on the traditional *Akaike Information Criterion* (AIC). The form of WAIC is similar to AIC, but it replace maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ by a log point-wise predictive density (lppd) and replaces k, the number of estimated parameters in AIC, with a data-based bias correction,

$$\text{WAIC} = 2p_{\text{WAIC}} - 2\sum_{i=i}^{n} \ln p_{\text{posterior}}(y_i) \tag{2.26}$$

where p_{WAIC} is defined as

$$p_{\text{WAIC}} = 2\sum_{i=1}^{n} \left\{ \ln \left[\mathbb{E}_{\text{posterior}} p(y_i|\boldsymbol{\theta}) \right] - \mathbb{E}_{\text{posterior}} \left[\ln p(y_i|\boldsymbol{\theta}) \right] \right\} \tag{2.27}$$

$$\approx 2\sum_{i=1}^{n} \left\{ \ln \left[\frac{1}{L} \sum_{\ell=1}^{L} p(y_i|\boldsymbol{\theta}^{(\ell)}) \right] - \frac{1}{L} \sum_{\ell=1}^{L} \ln p(y_i|\boldsymbol{\theta}^{(\ell)}) \right\} \tag{2.28}$$

LOO-CV Partition the data repeatedly into $\{\mathbf{y}_{\text{train}}, y_{\text{test}}\}$, then fit $\mathbf{y}_{\text{train}}$ and obtain lppd of y_{test}, for each sample y_i, the corresponding LOO-CV is

$$\begin{aligned}
\text{LOO-CV} &= \sum_{i=1}^{n} \ln p_{\text{posterior}(-i)}(y_i) \\
&= \sum_{i=1}^{n} \ln \int p(y_i|\boldsymbol{\theta}) p(\theta|\mathbf{y}_{-i}) d\boldsymbol{\theta} \\
&\approx \sum_{i=1}^{n} \ln \left\{ \frac{1}{L} \sum_{\ell=1}^{L} p(y_i|\boldsymbol{\theta}^{(\ell)}, \mathbf{y}_{-i}) \right\}
\end{aligned} \tag{2.29}$$

Gelman, Hwang, and Vehtari, 2014 also proposed and recommended an alternative to determine p_{WAIC} by computing posterior variance of the log density. They claimed that "its series expansion has closer resemblance to the series expansion for LOO-CV and also in practice seems to give results closer to LOO-CV".

Vehtari, Gelman, and Gabry, 2017 proposed an efficient computation of LOO-CV using *Pareto Smoothing Importance Sampling* (PSIS-LOO) and argued that PSIS-LOO "is more robust in the finite case with weak priors or influential observations". This is implemented in the `loo`[4] package.

2.6 Spatial Interpolation

In geostatistics, spatial interpolation is also referred as *kriging*, named by Matheron (1963) in honor of D.G. Krige, a South African mining engineer.

Let $Y(\mathbf{s}_{\text{obs}}) = \mathbf{y}_{\text{obs}}$ same as before, denote unknown value at new location as $Y(\mathbf{s}_{\text{new}}) = \mathbf{y}_{\text{new}}$ and the corresponding covariates of $(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{new}})$ as $\{\mathbf{x}(\mathbf{s}_{\text{obs}}), \mathbf{x}(\mathbf{s}_{\text{new}})\} = (\mathbf{x}_{\text{obs}}, \mathbf{x}_{\text{new}})$. Kriging under Bayesian framework is calculating the posterior predicted distribution.

$$\begin{aligned} p(\mathbf{y}_{\text{new}}|\mathbf{y}_{\text{obs}}, \mathbf{x}_{\text{obs}}, \mathbf{x}_{\text{new}}) &= \int p(\mathbf{y}_{\text{new}}, \boldsymbol{\Theta}|\mathbf{y}_{\text{obs}}, \mathbf{x}_{\text{obs}}, \mathbf{x}_{\text{new}}) d\boldsymbol{\Theta} \\ &= \int p(\mathbf{y}_{\text{new}}|\mathbf{y}_{\text{obs}}, \mathbf{x}_{\text{obs}}, \mathbf{x}_{\text{new}}, \boldsymbol{\Theta}) p(\boldsymbol{\Theta}|\mathbf{y}_{\text{obs}}, \mathbf{x}_{\text{obs}}) d\boldsymbol{\Theta} \end{aligned} \tag{2.30}$$

where $\boldsymbol{\Theta}$ is the parameter vector.

Kriging under the Bayesian framework, the uncertainty will also be described by the posterior predicted distribution credible intervals.

There are several ways to implement kriging. The basic idea is to the property that for a multivariate normal, the conditional distribution is still a multivariate normal.

Recall that the log scale of latent spatial structure is under the Gaussian process assumption. Since $Z(\mathbf{s}) \sim \text{GP}(\mathbf{0}, \boldsymbol{\Sigma})$, partitioning $Z(\mathbf{s}) = (\mathbf{z}_{\text{obs}}, \mathbf{z}_{\text{new}})^\top$, the dimension of $\mathbf{z}_{\text{obs}}$ is n_{obs} and for $\mathbf{z}_{\text{new}}$ is n_{new} respectively. For a multivariate normal,

$$\mathbf{z} = \begin{bmatrix} \mathbf{z}_{\text{obs}} \\ \mathbf{z}_{\text{new}} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_{\text{obs}} \\ \boldsymbol{\mu}_{\text{new}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\text{obs}} & \boldsymbol{A} \\ \boldsymbol{A}^\top & \boldsymbol{\Sigma}_{\text{new}} \end{bmatrix} \right)$$

with $\boldsymbol{\Sigma}_{\text{obs}}$ be a $n_{\text{obs}} \times n_{\text{obs}}$, $\boldsymbol{\Sigma}_{\text{new}}$ be a $n_{\text{new}} \times n_{\text{new}}$ covariance matrix of the observed and new locations respectively. $\boldsymbol{A}$ is a $n_{\text{obs}} \times n_{\text{new}}$ covariance matrix between observed and new locations and $\boldsymbol{A}^\top$ is of the dimension $n_{\text{new}} \times n_{\text{obs}}$.

[4] http://mc-stan.org/loo/

The conditional distribution of $\mathbf{z}_{\text{new}}|\mathbf{z}_{\text{obs}} \sim N\left(\boldsymbol{\mu}_{\text{new}|\text{obs}}, \boldsymbol{\Sigma}_{\text{new}|\text{obs}}\right)$, where

$$\begin{aligned}\boldsymbol{\mu}_{\text{new}|\text{obs}} &= \boldsymbol{\mu}_{\text{new}} + \boldsymbol{A}^\top \boldsymbol{\Sigma}_{\text{obs}}^{-1}(\mathbf{z}_{\text{obs}} - \boldsymbol{\mu}_{\text{obs}}) \\ \boldsymbol{\Sigma}_{\text{new}|\text{obs}} &= \boldsymbol{\Sigma}_{\text{new}} - \boldsymbol{A}^\top \boldsymbol{\Sigma}_{\text{obs}}^{-1}\boldsymbol{A}\end{aligned}$$

In our case, $\boldsymbol{\mu}_{\text{new}}$ and $\boldsymbol{\mu}_{\text{obs}}$ are both assumed zero vectors, the equations become

$$\boldsymbol{\mu}_{\text{new}|\text{obs}} = \boldsymbol{A}^\top \boldsymbol{\Sigma}_{\text{obs}}^{-1}\mathbf{z}_{\text{obs}} \tag{2.31}$$

$$\boldsymbol{\Sigma}_{\text{new}|\text{obs}} = \boldsymbol{\Sigma}_{\text{new}} - \boldsymbol{A}^\top \boldsymbol{\Sigma}_{\text{obs}}^{-1}\boldsymbol{A} \tag{2.32}$$

The algorithm of Kriging for the new locations is as follows,

Algorithm 4: Kriging for Gamma spatial model. Appendix (A.2)

Input: Posterior parameters $\{\boldsymbol{\theta}, \beta, \sigma^2, \phi, \mathbf{z}_{\text{obs}}\}$ and all locations $\{\mathbf{s}_{\text{obs}}, \mathbf{s}_{\text{new}}\}$
Output: Predicted values $\mathbf{y}_{\text{new}}$

```
Compute distance matrix D from {s_obs, s_new}
Partition D into 4 blocks D_obs, D_new, D_A, D_A^⊤
for 1 to n do
    Compute Σ_obs = σ^2 exp{−D_obs/φ}
    Compute Σ_new = σ^2 exp{−D_new/φ}
    Compute A = σ^2 exp{−D_A/φ} and transpose to get A^⊤
    Compute Σ_obs^{-1}                                   O(n_obs^3)
    Compute μ_new|obs = A^⊤ Σ_obs^{-1} z_obs             O(n_new n_obs^3)
    Compute Σ_new|obs = Σ_new − A^⊤ Σ_obs^{-1} A         O(n_new^2 n_obs^3)
    Simulate z_new ~ N(μ_new|obs, Σ_new|obs)             O(n_new^3)
    Compute α(s_new) = exp(s^⊤ θ + z_new)
    Simulate y_new ~ Gamma(β · α(s_new), β)
```

By Algorithm 4, we handle the partition of multivariate normal outside MCMC.

Alternatively, we may treat $\mathbf{y}_{\text{new}}$ as missing values, then construct the model as Equation (2.10), by this way, the partition of multivariate normal is embedded in MCMC. Note that if we perform kriging as missing values through `Stan`. Then all procedures are handled internally by `Stan`.

It is preferable to implement kriging through Algorithm 4. Computationally, within each iteration, the main intensive task in HMC is to evaluate the gradient for log-posterior kernel by leapfrog approximation of Hamiltonian system, the time complexity is roughly $\mathcal{O}(\mathcal{L}(n_{\text{obs}} + n_{\text{new}})^2)$ where recall that $\mathcal{L}$ is number of leapfrog steps. For NUTS algorithm, $\mathcal{L}$ is not fixed, large scale of data corresponds to a high dimensional posterior kernel, $\mathcal{L}$ may go up to the scale of $n_{\text{obs}} + n_{\text{new}}$. Moreover, sampling $\mathbf{z}_{\text{new}}$, involves a Cholesky decomposition of covariance matrix which has time complexity

$\mathcal{O}((n_{\text{obs}} + n_{\text{new}})^3)$. On the other hand, in Algorithm 4, MCMC only deal with the matrix of dimension n_{obs}, with time complexity around $\mathcal{O}(n_{\text{obs}}^3)$. But it will still take some time to obtain kriging samples after getting getting parameters. The difference time of these two methods become notable when the number of new locations increases.

It may be worth mentioning that for Algorithm 4. Sampling steps can also done in `Stan`. This suppose to be faster but not recommended by the author. `Stan` stores every parameters it used in the memory, sampling with `Stan` may lead to space complexity increases drastically, this may eventually result in slower access to data. Algorithm 4 can be further improved by using lower level programming language such as `C++` and parallel computing. Figure 2.4 compare an experiment of kriging in time for different number of new location by fitting a gamma spatial model.

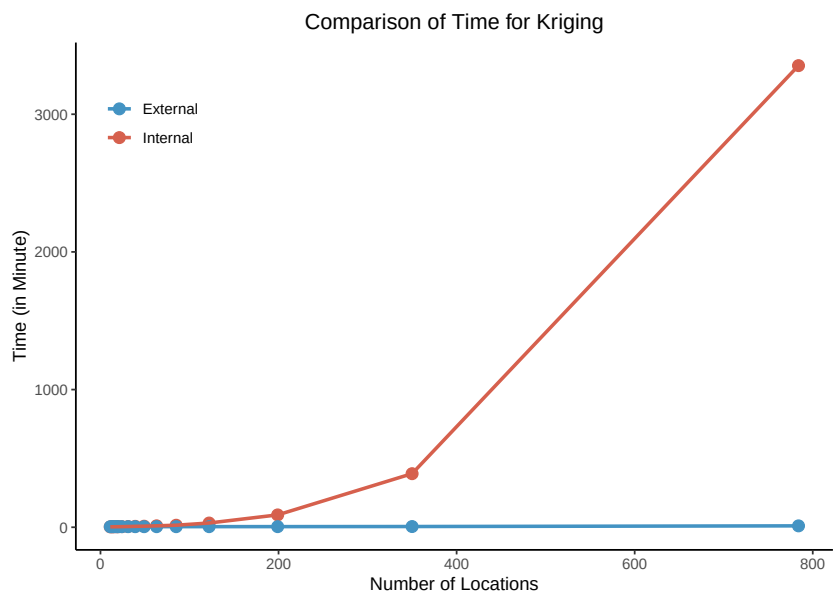


FIGURE 2.4: Comparison in time (in minute) between kriging implementations against number of locations. "•" represents kriging as missing values. "•" stands for interpolation through Algorithm 4. The MCMC fits a gamma spatial models with 3 Markov chains, each chain runs 1500 iterations, 750 iterations are warm-up samples.

Chapter 3

Data Analysis

In this chapter, we will first perform a simulation study to investigate the performance of different setting of the prior specifications for the scale parameters $\{\beta, \sigma^2\}$, then we choose the best setting for the model to fit Comilla's under different models and compare their performance. We then will predict the As level of the whole region of Comilla. Finally, we will compare our the performance of our gamma model to other models that are also able to handle skew processes.

3.1 Simulation Study

As it is mentioned in Section 2.4.2, it important to investigate a proper choice of the prior distribution for the scale parameters $\{\beta, \sigma^2\}$ since they have direct contribution to the marginal mean and variance under gamma model.

In this simulation study, we will use the location information from Comilla. There are 4 locations that have more than one observations, we know from the BGS description that they are different tube wells. We randomly move overlap locations by a small distance (less than 1 km) to avoid computation issue. The jitter locations ensure that the positive definite condition of covariance matrix holds. We assign our "true parameters" to generate random samples follow the hierarchical model describe in Equation (2.10). Then by fitting the same spatial model, with different specifications, we are mainly going to investigate

1. If we are able to recover the parameters.
2. How the posterior distribution of the scale parameters behave under different prior specifications.
3. What is the performance of the predicted values.

We use *Root Mean Square Error* (RMSE) to measure the performance of the predicted values since we "known the truth".

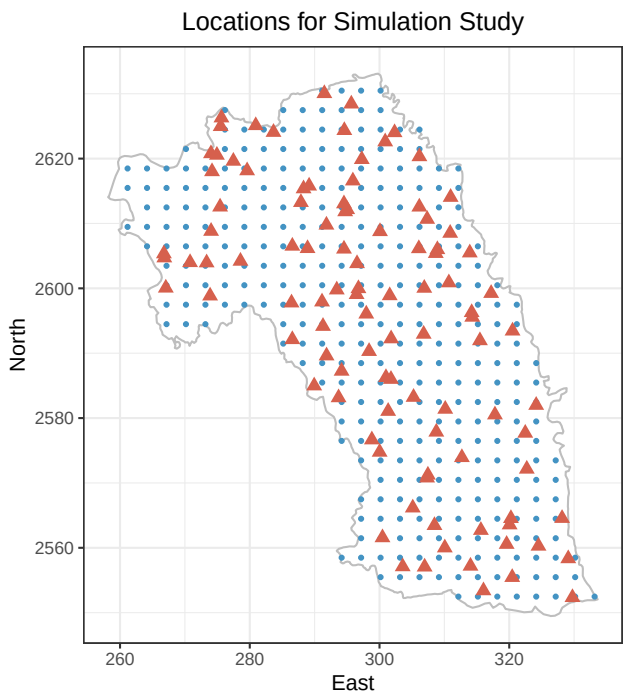


FIGURE 3.1: Locations in Comilla used for simulation study, "▲" denotes observed locations ($n_{\text{obs}} = 101$); "●" are new locations to be predicted ($n_{\text{new}} = 350$).

Figure 3.1 illustrates the locations used in the study, Table 3.1 summarizes the information of the location. The locations is used to determine the parameter of prior distribution for ϕ.

	Min	Max
distance (km)	0.315	91.629

TABLE 3.1: Summary of the distance between locations in Comilla

We assume the log-mean of the gamma distribution to be described by a linear combination of the easting and northing coordinates. The values chosen for the parameters of the model are shown in Table 3.2. We elaborate steps to simulate random samples,

1. Choose and fixed the "true" parameters $\{\boldsymbol{\theta}, \beta, \sigma^2, \phi\}$

Parameter	θ_0	θ_1	θ_2	β	σ^2	ϕ
Value	4	1	-0.5	0.1	0.5	18

TABLE 3.2: True Parameters for Simulations

2. Compute pairwise distance of all locations $\mathbf{D}$.

3. Compute $\mathbf{\Sigma}$ through exponential correlation function according to $\{\sigma^2, \phi\}$.

4. Simulate $Z(\mathbf{s}) = \mathbf{z} = (\mathbf{z}_{\text{obs}}, \mathbf{z}_{\text{new}})^\top \sim N(\mathbf{0}, \mathbf{\Sigma})$.

5. Compute $\alpha(\mathbf{s}) = \exp\left[\theta_0 + \theta_1 \cdot \text{east}(\mathbf{s}) + \theta_2 \cdot \text{north}(\mathbf{s}) + Z(\mathbf{s})\right]$ according to $\{\boldsymbol{\theta}, \mathbf{s}\}$

6. Simulate $\mathbf{y}^{(\ell)} = (\mathbf{y}^{(\ell)}_{\text{obs}}, \mathbf{y}^{(\ell)}_{\text{new}})^\top \sim \text{Gamma}(\beta \cdot \alpha(\mathbf{s}), \beta)$ according to $\{\beta\}$ under different seeds, where seed $\ell \in \{1, \ldots, 50\}$.

We now have 50 different samples of size $n = n_{\text{obs}} + n_{\text{new}} = 451$. We proceed with model fitting. Prior distributions for $\{\boldsymbol{\theta}, \phi\}$ will follow Section 2.4.2, where $\theta_i \sim N(0, 10^2), i = 0, 1, 2$ and $\phi \sim \text{InvGamma}(2, 16)$. We start from fixing priors for $\sigma \sim \text{InvGamma}(2, 1)$ since we know the true value of $\sigma^2 = 0.5 (\sigma \approx 0.7)$. Assigning priors to σ instead of σ^2 is suggested by Gelman, 2006. This prior distribution will properly cover the range of σ^2 and provide some information about the "truth".

3.1.1 Scale Parameter for Global Process: β

For β we will compare three different candidates distributions,

P1 half-Cauchy: $\mathcal{HC}(2)$;

P2 Uniform: $U(0, 0.3)$;

P3 Inverse Gamma: $\text{InvGamma}(1, 1)$.

The first two priors are proposed by Gelman, 2006. We can take large value for scale parameter to half-Cauchy. For uniform prior, in general, we should assign a large value for the upper bound. In this study, however, since we know the "truth", we may try to limit the upper bound to see the performance. The $\text{InvGamma}(1, 1)$ is a strong prior with mode away from the "truth", we want to see by assigning "False" prior information, how the posteriors will behave. Figure 3.2 illustrates $\mathcal{HC}(2)$ and $\text{InvGamma}(1, 1)$.

We continue describing the fitting and examining procedures. For each sample,

1. Fit P1 to P3 on $\mathbf{y}^{(\ell)}_{\text{obs}}$ only.

2. Obtain posterior median and 95% credible interval of $\{\beta, \sigma^2\}$ for P1 to P3.

3. Obtain posterior median $\hat{\mathbf{y}}^{(\ell)}_{\text{new}}$ through Algorithm 4, for P1 to P3.

4. Compute $\text{RMSE}^{(\ell)} = \sqrt{n_{\text{new}}^{-1} \sum_\ell (\mathbf{y}^{(\ell)}_{\text{new}} - \hat{\mathbf{y}}^{(\ell)}_{\text{new}})^2}$ for P1 to P3.

Figure 3.3 illustrates the result of 95% credible interval of β for 50 different samples under P1 to P3.

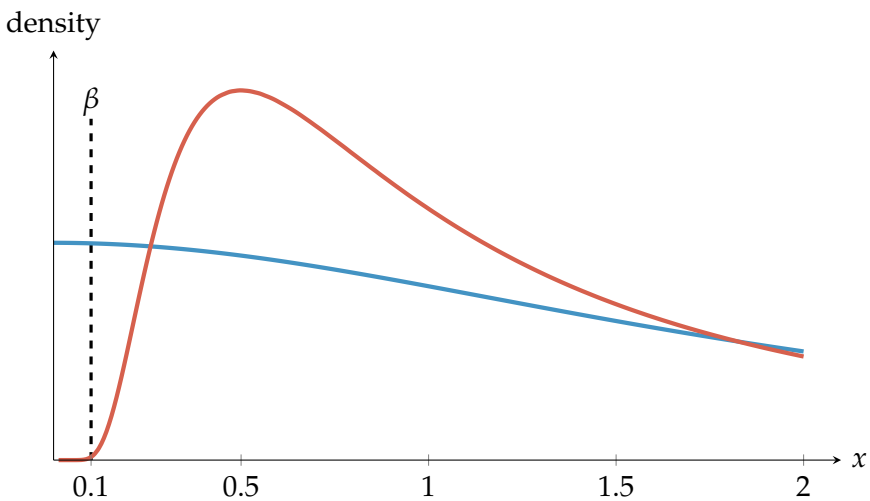


FIGURE 3.2: Density for $\mathcal{HC}(2)$ and $\text{InvGamma}(1,1)$, black dashed line is the "true" value of parameter β. half-Cauchy is weakly informative prior and Inverse Gamma with this parameter setting is away from the "truth".

P1 $\mathcal{HC}(2)$ For most of the time, the posterior credible interval covers the true value of β. For samples $\ell \in \{16, 19, 23, 24\}$ the variation the credible interval is large. Note that the prior for $\mathcal{HC}(2)$ is quite vague, but the posterior credible interval are quite narrow. This may suggest that by this model specification, the scale parameter β may carry enough information to be recovered.

P2 $U(0, 0.3)$ Because we limit the upper bound, the posterior credible interval cannot surpass 0.3. But it may be notable that except for samples $\ell \in \{16, 19, 23, 24\}$, for other samples, the posterior credible interval of uniform has similar scale as half-Cauchy.

P3 $\text{InvGamma}(1, 1)$ Because the prior information is so biased that even posterior parameter have learn much information may not enough for the truth to fall into the 95% credible interval; The large variation of the credible interval for β is also true under the same samples $\ell \in \{16, 19, 23, 24\}$.

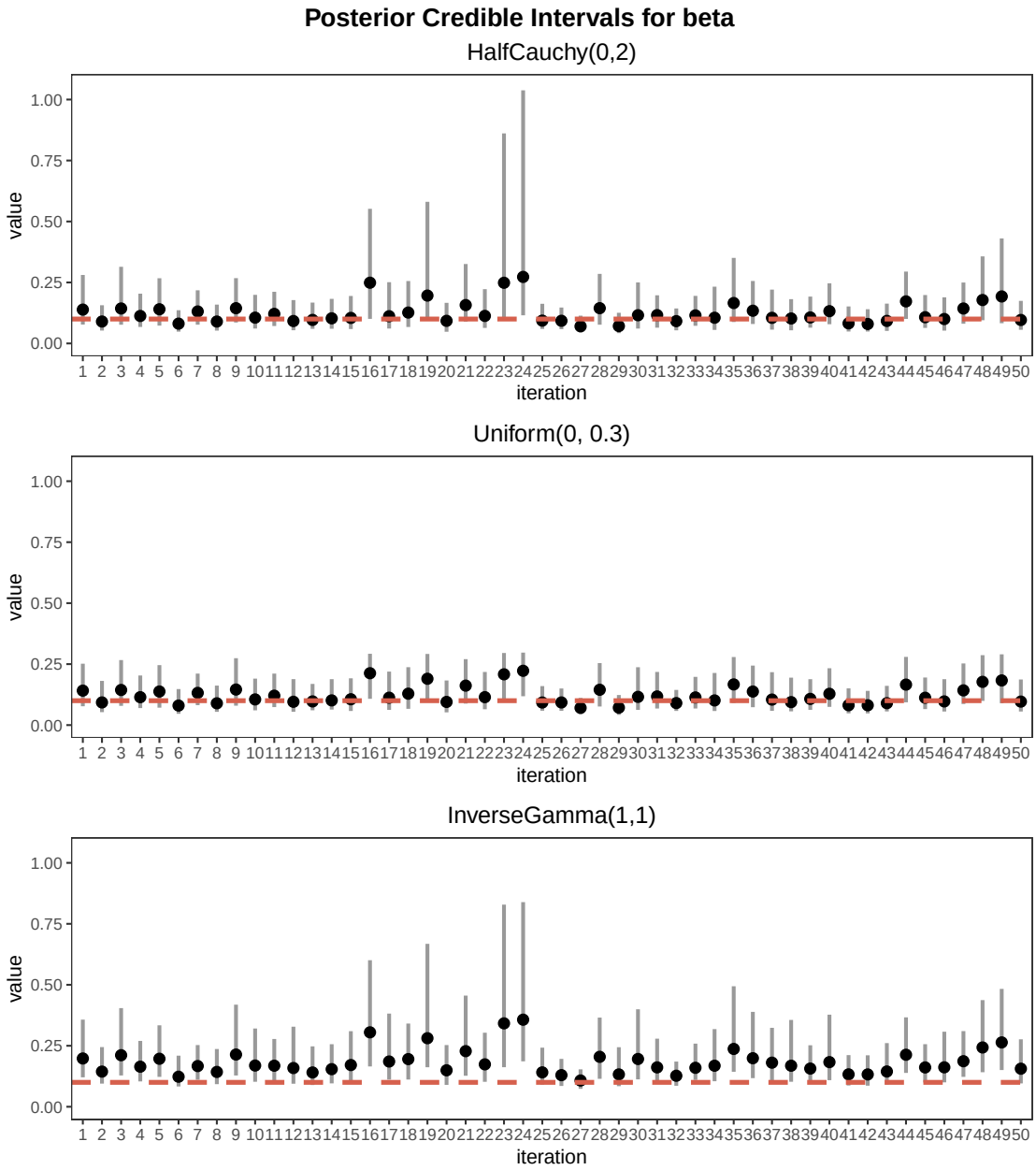


FIGURE 3.3: Posterior Credible Interval for β under different priors for β. P1: $\mathcal{HC}(2)$; P2: $U(0, 0.3)$ and P3: InvGamma$(1, 1)$. "•" is the posterior median, the gray vertical line is the 95% credible interval, red dashed lines are true values of the parameter.

Figure 3.4 illustrates the result of 95% credible interval of σ^2 for 50 different samples under P1 to P3. It seems that all three priors will lead to σ^2 being underestimated against true values, this is logical since P1 to P3 are assigned same priors for σ^2. Also note that the behavior of σ^2 is similar under different samples for P1 to P3.

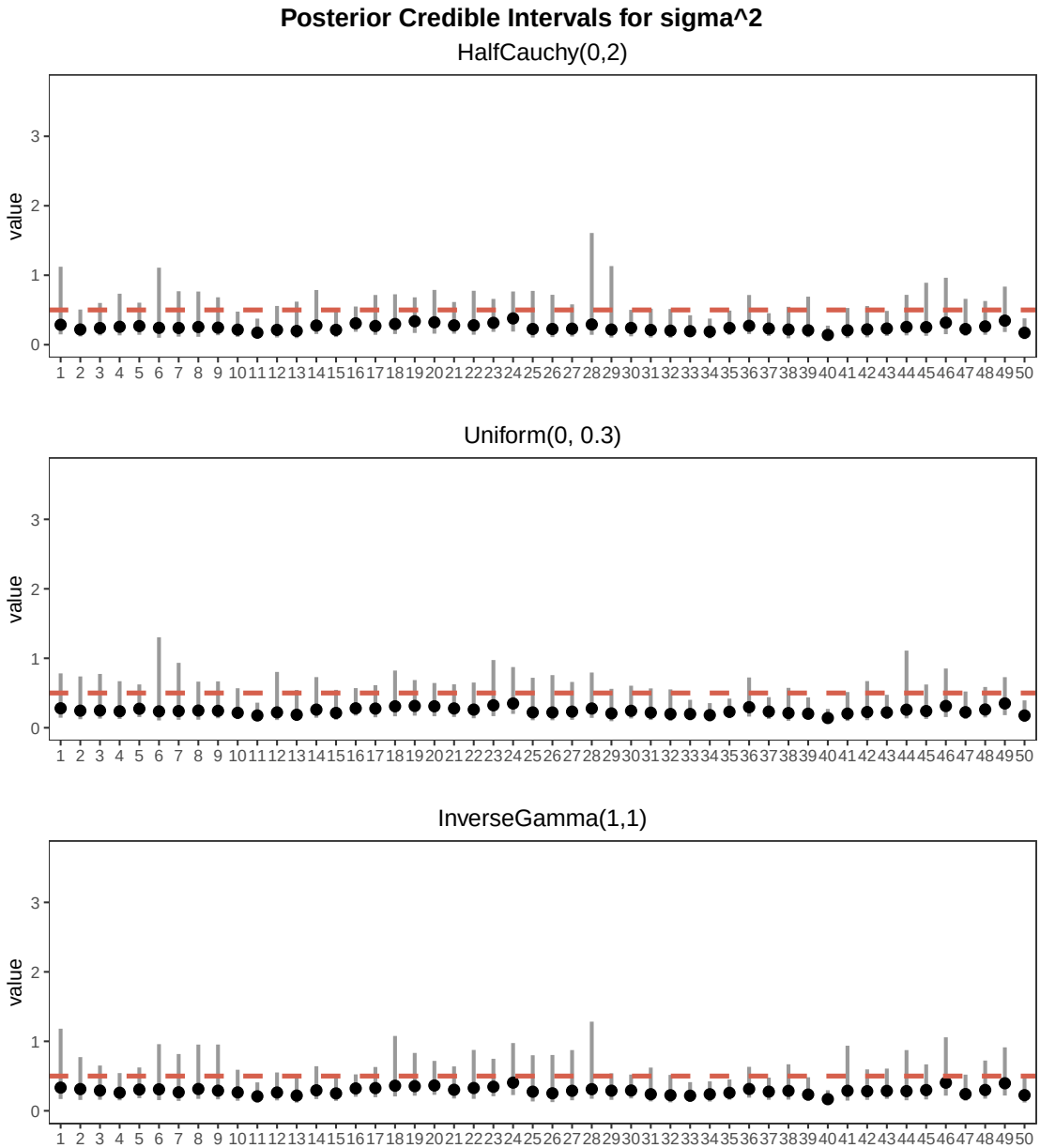


FIGURE 3.4: Posterior Credible Interval for σ^2 under different priors for β. P1: $\mathcal{HC}(2)$; P2: $U(0, 0.3)$ and P3: InvGamma$(1, 1)$. "•" is the posterior median, the gray vertical line is the 95% credible interval, red dashed lines are true values of the parameter.

	$\mathcal{HC}(2)$	$U(0, 0.3)$	InvGamma$(1, 1)$
RMSE	119.190	119.410	120.506

TABLE 3.3: Mean of RMSE of posterior predicted median for 50 iterations under different prior distributions for β

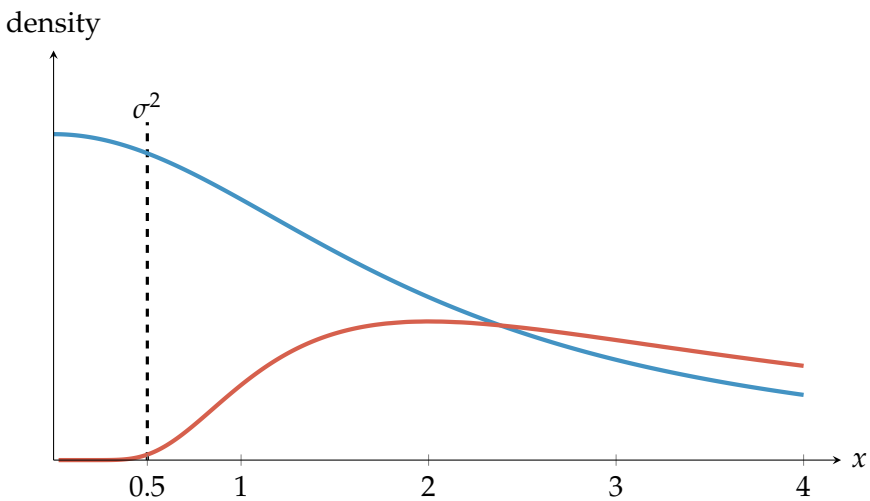


FIGURE 3.5: Density for $\mathcal{HC}(2)$ and InvGamma$(1, 4)$, black dashed line is the "true" value of parameter σ^2.

Table 3.3 shows the mean of RMSE for posterior predicted median of $\mathbf{y}_{\text{new}}$ of 50 samples. At this scale, they are quite close. $\mathcal{HC}(2)$ has the smallest mean of RMSE. This may suggest it would be appropriate to choose $\beta \sim \mathcal{HC}(2)$ as prior distribution, the choice of this prior is consistent with the suggestion from Polson and Scott, 2012.

3.1.2 Scale Parameter for Latent Spatial Process: σ^2

For σ^2, we follow same strategy, with $\beta \sim \mathcal{HC}(2)$ fixed, we compare the following priors for σ for comparison. Figure 3.5 illustrates $\mathcal{HC}(2)$ and InvGamma$(1, 4)$.

P4 half-Cauchy: $\mathcal{HC}(2)$;

P5 Uniform: $U(0, 2)$;

P6 Inverse Gamma: InvGamma$(1, 4)$.

Figure 3.6 illustrates the result of 95% credible interval of β for 50 different samples under P4 to P6.

We can see that posterior behavior for β is similar for P4 to P6, and it seems that different prior specifications for σ^2 has no major impact for the posterior on β.

Figure 3.7 shows the result for 95% credible interval of σ^2 for 50 different samples under P4 to P6.

Underestimate Although the "true" values of σ^2 are covered by most the 95% credible intervals for different samples, all three priors still cannot avoid underestimate from the "truth".

Comparison among P4 to P6 Compare to three different priors, when assigning $\mathcal{HC}(2)$, the variation for σ^2 seems less volatile than $U(0, 2)$ and $\text{InvGamma}(1, 4)$.

Comparison among P1 to P6 However, note that the widest credible is produced under $\mathcal{HC}(2)$ at sample seed $\ell = 18$, the upper limit goes up to 3.669. This already exceeds all of the upper limit when fixing $\sigma \sim \text{InvGamma}(2, 1)$.

Scale of σ^2 It worth notice that the scale of σ^2 under P1 to P3 shown in Figure 3.4 (from 0 to 1.65) and under P4 to P6 shown in Figure 3.7 (from 0 to 3.7) and are in different scales. These may suggest that vague priors (P4 and P5) or misleading prior (P6) may result the estimate unstable.

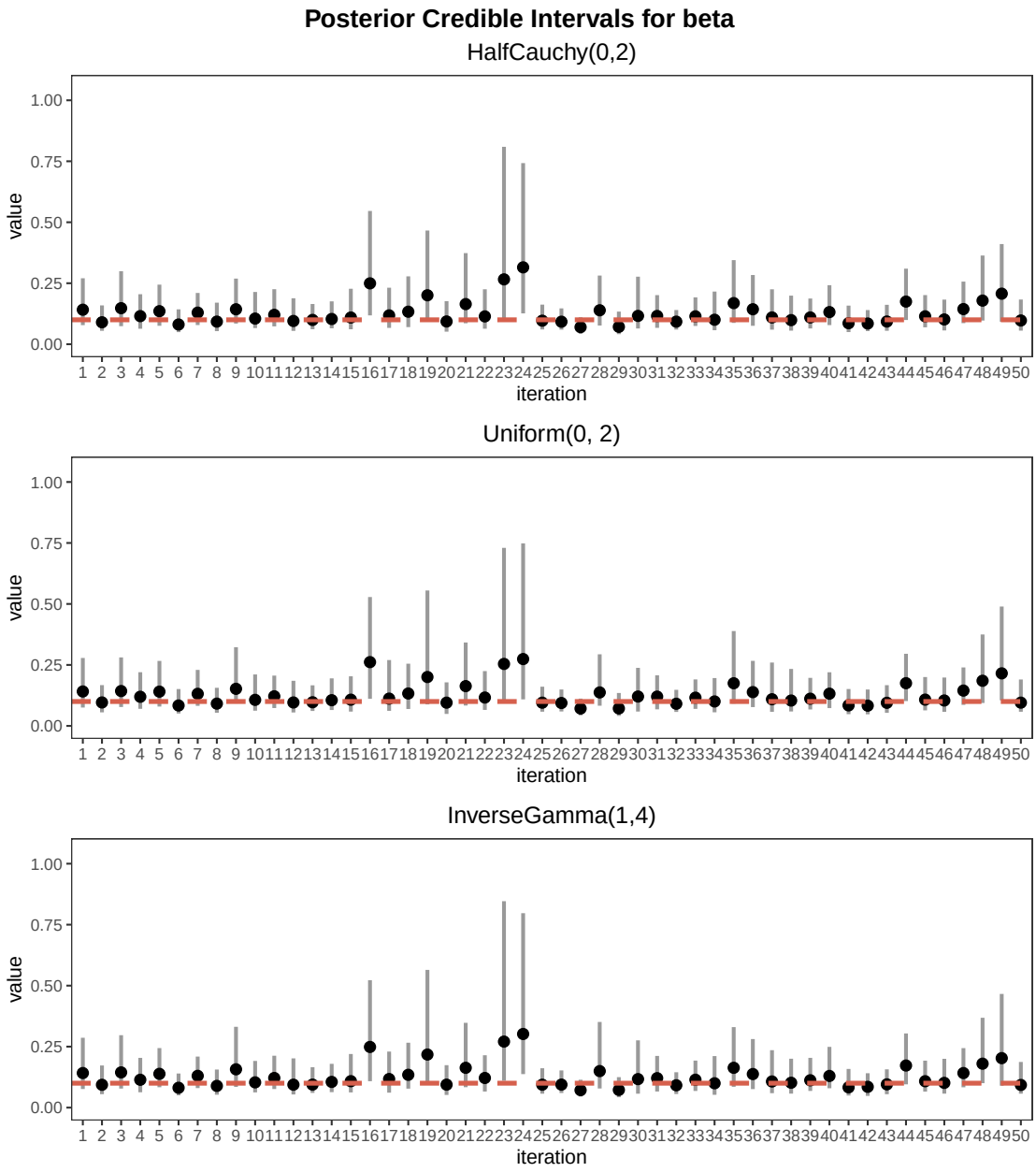


FIGURE 3.6: Posterior Credible Interval for β under different priors for σ^2. P4: $\mathcal{HC}(2)$; P5: $U(0,2)$ and P6: InvGamma$(1,4)$. "•" is the posterior median, the gray vertical line is the 95% credible interval, red dashed lines are true values of the parameter.

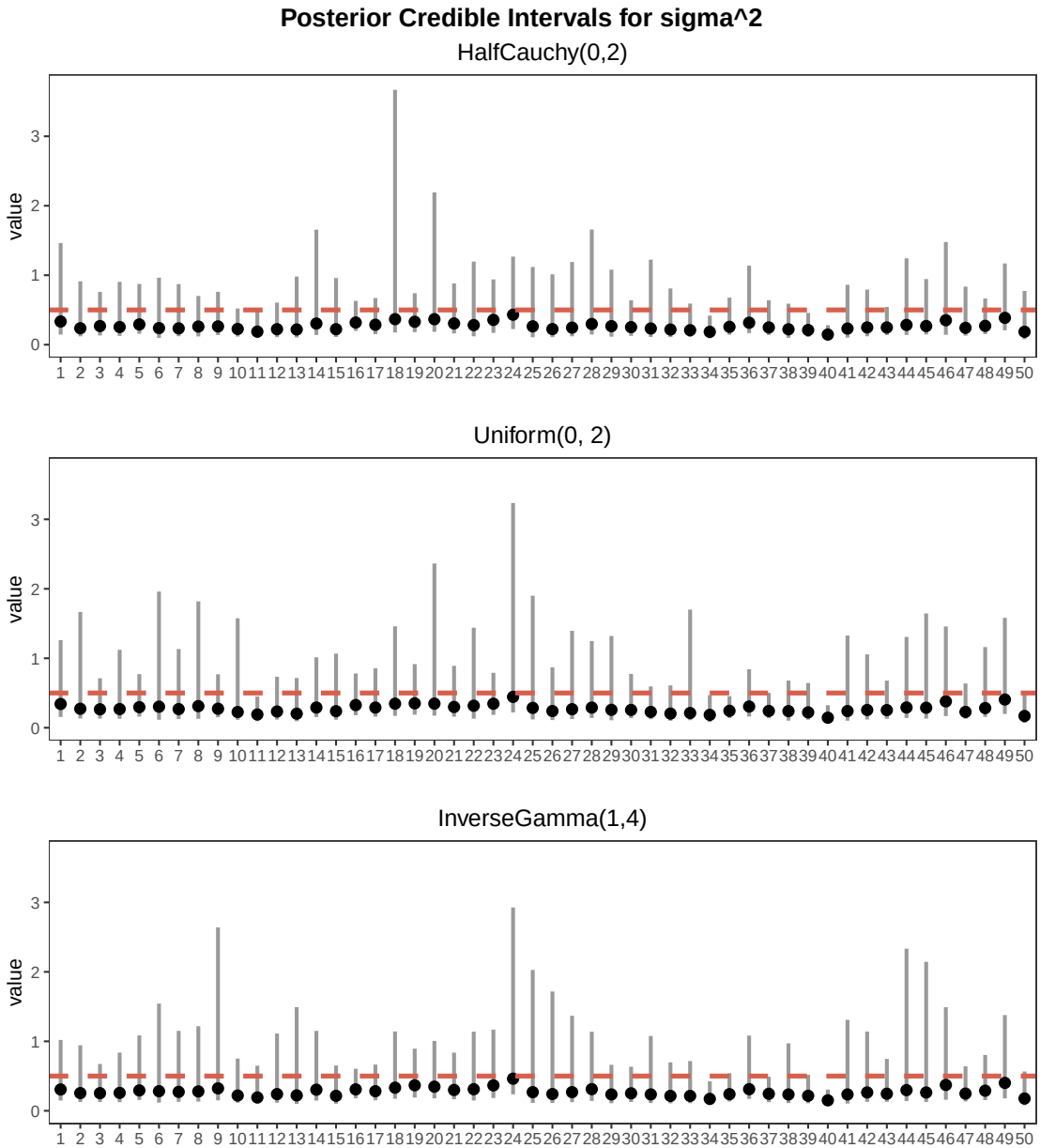


FIGURE 3.7: Posterior Credible Interval for σ^2 under different priors for σ^2. P4: $\mathcal{HC}(2)$; P5: $U(0,2)$ and P6: InvGamma$(1,4)$. "•" is the posterior median, the gray vertical line is the 95% credible interval, red dashed lines are true values of the parameter.

Table 3.4 summarizes the result for the mean of RMSE, $\mathcal{HC}(2)$ has the best performance, but taking Table 3.3 into account, we believe that P1 is desired among all simulations. We will use the prior setting of P1 for the real Comilla data.

	$\mathcal{HC}(2)$	$U(0,2)$	InvGamma$(1,4)$
RMSE	119.216	119.388	119.238

TABLE 3.4: Mean of RMSE of posterior predicted median for 50 iterations under different prior specifications for σ^2

3.2 Arsenic in Comilla

The data set is publicly available from the BGS website [1], it was released on May 25th 2000.

Comilla's data contains a sample of size $n = 110$ tubewells, 9 of which are censored data, with 4 coded as " < 6 " and the other 5 recorded as " < 0.5 ". Each sample contains 34 variables, 14 of them are descriptive information and 20 are measurements.

Descriptive: `sample_id, sample_field_id, sample_date, lat_deg, long_deg, year_construction, well_type, well_depth, division, district, district, thana, union, mouza, geocode`.

The locations of nearly all sample sites were established by hand-held Global Positioning System (GPS) devices which at the time of sampling (1998/99) were accurate to within about 50 –100 m. (Kinniburgh and Smedley, 2001)

Measurement The arsenic concentration level (denoted as `As`) is recorded by $\mu g/L$. Other 19 chemical concentrations are listed using mg/L. A list of the measurement is as follows (denoted by their chemical symbols), `As, Al, B, Ba, Ca, Co, Cr, Cu, Fe, K, Li, Mg, Mn, Na, P, Si, SO4, Sr, V, Zn`. We are only interested in the As levels.

The map of Comilla is obtained from the Global Administrative Areas (GADM) website. We choose version 2.8 for mapping [2] (the latest version is 3.6) and choose the map level detail up to district level (since Comilla is a district). The coordinate reference system is longitude/latitude and the WGS84 datum, the correspond UTM Zone for Comilla is 46N.

3.2.1 Exploratory Data Analysis

Figure 3.9 shows the location and type of observations for Comilla's data. There are $n_{\text{obs}} = 101$ observations and $n_{\text{cen}} = 9$ left censored data, 4 of which are coded as " < 6 " and the rest 5 are recorded as " < 0.5 ".

[1] http://www.bgs.ac.uk/research/groundwater/health/arsenic/Bangladesh/data.html

[2] https://gadm.org/download_country_v2.html

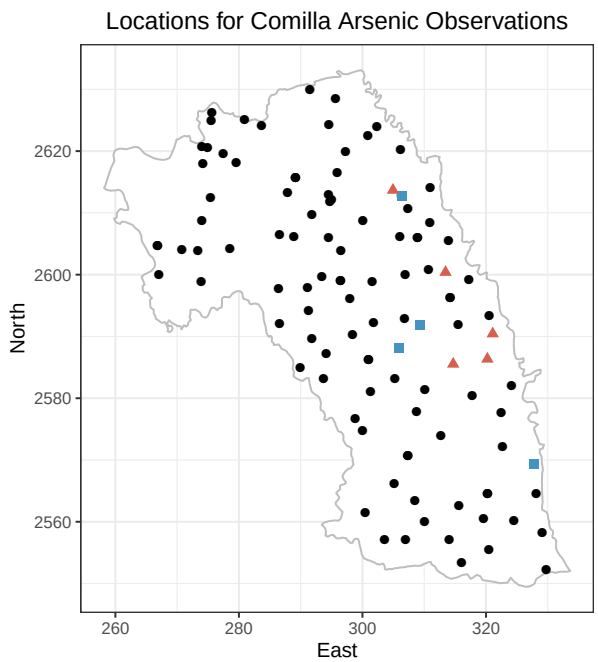


FIGURE 3.8: Location information for Comilla arsenic data. "•" represent observations, "▲" is censored data labeled as " < 0.5", "■" is censored data labeled as " < 6".

Some EDA help us to obtain a general look from the arsenic in Comilla. We can see from Table 3.5 that As level spread from 0.5 to 698.00 and left hand side of Figure 3.9 reveals that the As level at Comilla is right skewed. The right hand side of Figure 3.9 suggests a clue of existing spatial structures of As in Comilla.

	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Arsenic ($\mu g/L$)	0.50	6.00	110.00	141.76	211.75	698.00

TABLE 3.5: Summary of the arsenic levels

Further, in Figure 3.10, the left hand side shows that there exists some relations between the northing and As levels, and the graph on right hand side implies that the relation seems stronger on easting against As.

3.2.2 Model Fitting and Comparison

We fit models according to Table 3.6. In practice, we standardize the covariates (easting and northing) to make them in similar scales, it may help the MCMC have better behavior. Here we only take easting and northing as covariates, because when performing kriging, these two are the only covariates that are easy to obtain. If covariates are

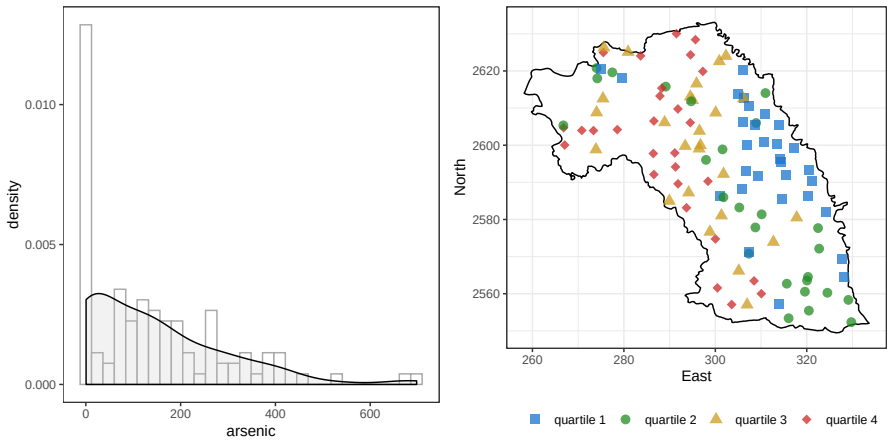


FIGURE 3.9: Histogram and Quartile plot of the Comilla arsenic levels.

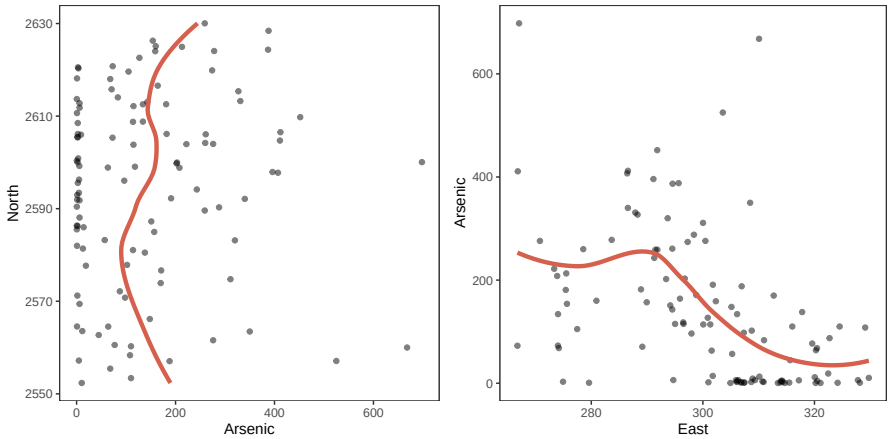


FIGURE 3.10: Easting and Northing against the arsenic levels.

obtained through prediction, the error may be magnified in the end. Table 3.7 lists the priors specifications, which follows the discussion in Section 3.1.

	Mean Structure $\mathbf{x}(\mathbf{s}_i)^\top \boldsymbol{\theta}$	Spatial Structure $Z(\mathbf{s})$
Model 1	θ_0	$\mathbf{0}$
Model 2	$\theta_0 + \theta_1 \cdot \text{east}(\mathbf{s}_i) + \theta_2 \cdot \text{north}(\mathbf{s}_i)$	$\mathbf{0}$
Model 3	θ_0	$\text{GP}(\mathbf{0}, \boldsymbol{\Sigma})$
Model 4	$\theta_0 + \theta_1 \cdot \text{east}(\mathbf{s}_i) + \theta_2 \cdot \text{north}(\mathbf{s}_i)$	$\text{GP}(\mathbf{0}, \boldsymbol{\Sigma})$

TABLE 3.6: Model Specifications for conditional log-mean of gamma models. Model 1 and 3 assume constant mean; Model 2 and 4 adding easting and northing as covariates; Model 3 and 4 adding Gaussian process as spatial structure.

For MCMCs, we use 3 Markov chains for each model. For each chain, we run 1500

Parameter	Distribution
$\theta_i, i = 0, 1, 2$	$N(0, 10^2)$
β	$\mathcal{HC}(2)$
σ	$\text{InvGamma}(2, 1)$
ϕ	$\text{InvGamma}(2, 16)$

TABLE 3.7: Prior specifications of gamma model for Comilla data

iterations with the first 750 set as burn-in period. Therefore, we will have 2250 posterior samples for all 3 chains. Convergence are checked for all models via trace-plot and $\hat{R}$, see Appendix B for example.

Table 3.8 summarizes the posterior distribution of the respective parameters under Models 1 to 4. Recall from Section 2.5, WAIC and LOO-CV are estimating point-wise out-of-sample prediction accuracy using point-wise log predictive density, we can only compute the $n_{\text{obs}} = 101$ observations.

	Model 1	Model 2	Model 3	Model 4
θ_0	4.941 (4.675, 5.220)	4.839(4.623, 5.078)	4.673(2.605, 5.784)	4.638 (3.705, 5.697)
θ_1		-0.820(-1.028, -0.613)		-0.982(-1.582, -0.354)
θ_2		-0.530(-0.764, -0.289)		-0.451(-1.069, 0.171)
β	0.003 (0.002, 0.005)	0.006(0.004, 0.008)	0.012(0.008, 0.020)	0.013(0.008, 0.021)
σ^2			1.165(0.559, 3.240)	0.737(0.343, 2.037)
ϕ			26.650(11.141, 87.265)	16.301(6.520, 60.335)
WAIC	1205.132	1175.066	1133.943	1137.110
LOO-CV	1205.141	1175.144	1140.386	1143.965

TABLE 3.8: Posterior Parameter Summary of Model 1 to Model 4 with posterior median (2.5%, 97.5%) percentiles, WAIC and LOO-CV

From Table 3.8, we see that 0 is not included in the 95% credible interval for θ_1, this implies that easting is a significant co-variate. For the parameter associated with northing, θ_2, 0 is not inside the 95% credible interval for Model 2 but is within the interval of Model 4. This suggests that the spatial structure is able to capture the information in south-north direction. Figure 3.11 shows the posterior distribution of $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)^\top$ for Model 4. The plots reveal that although the prior distribution is vague ($\theta_i \sim N(0, 10^2)$), the data have carried enough information.

For β, the global scale parameter, Figure 3.12 compares posterior distributions of β among different models. Along with Table 3.8, we can see that imposing spatial structures help increasing the values, which implies the spatial processes carry information about the variation of As level. This is similar to what we discussed earlier in Section 3.1.

On $\{\sigma^2, \phi\}$, Table 3.8 reveals that there seems to have some correlation, Model 3 has higher values for both $\{\sigma^2, \phi\}$ than Model 4. Figure 3.13 illustrates the posterior information about $\{\sigma^2, \phi\}$ and supports the idea of what we discussed.

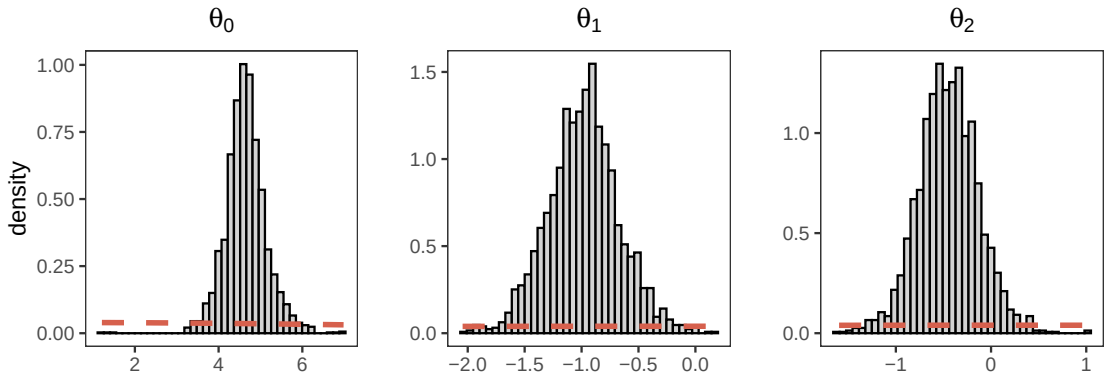


FIGURE 3.11: Posterior distributions of $\theta_i, i = 0, 1, 2$ for Model 4, red dashed lines are the prior distributions.

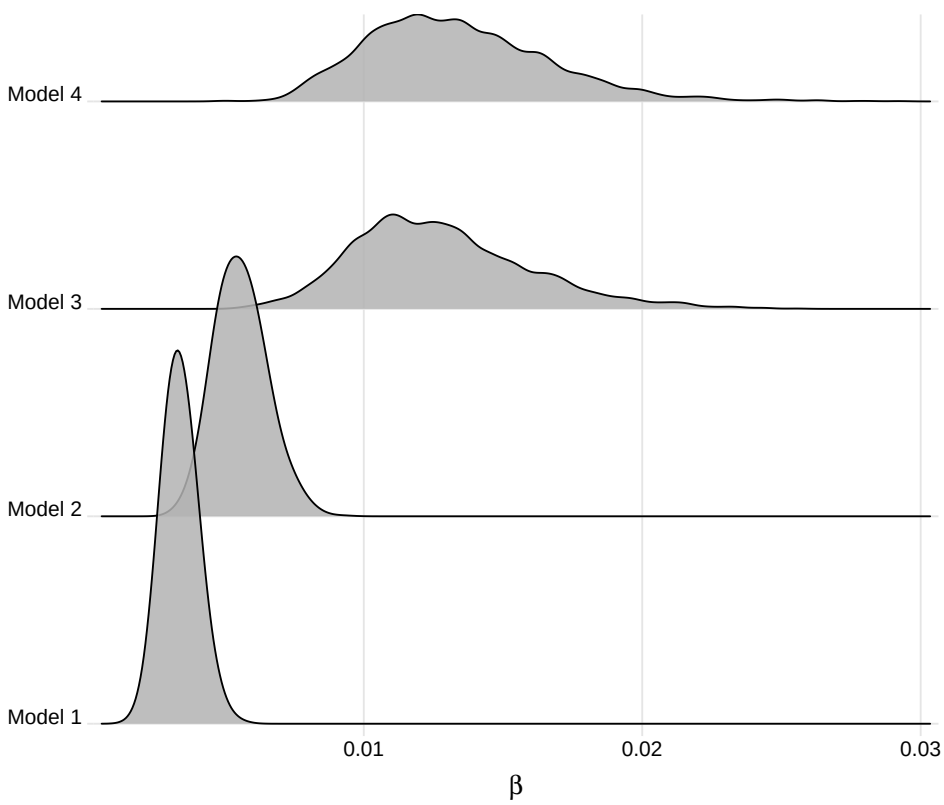
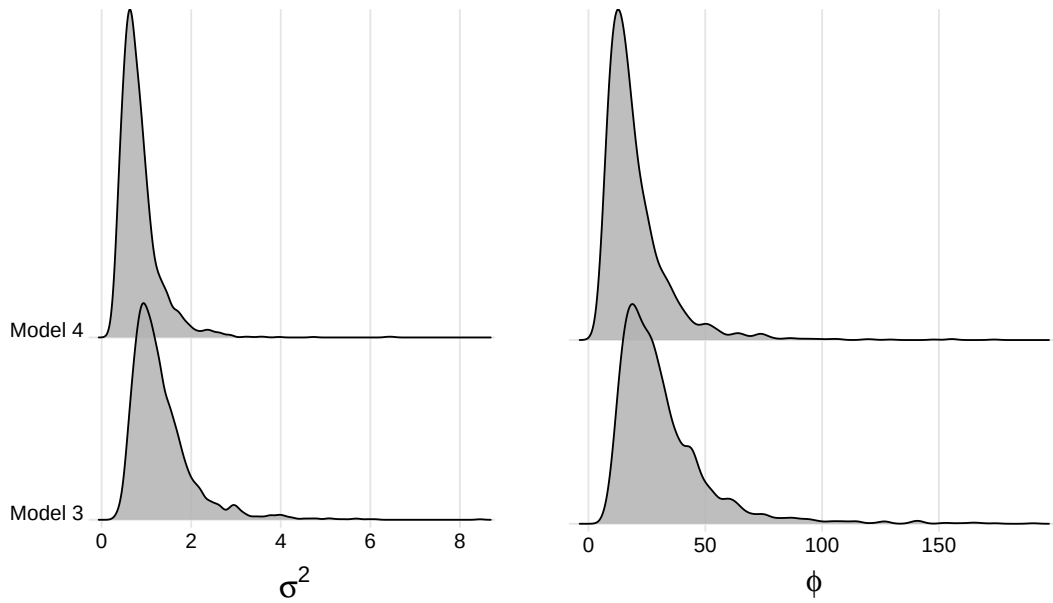
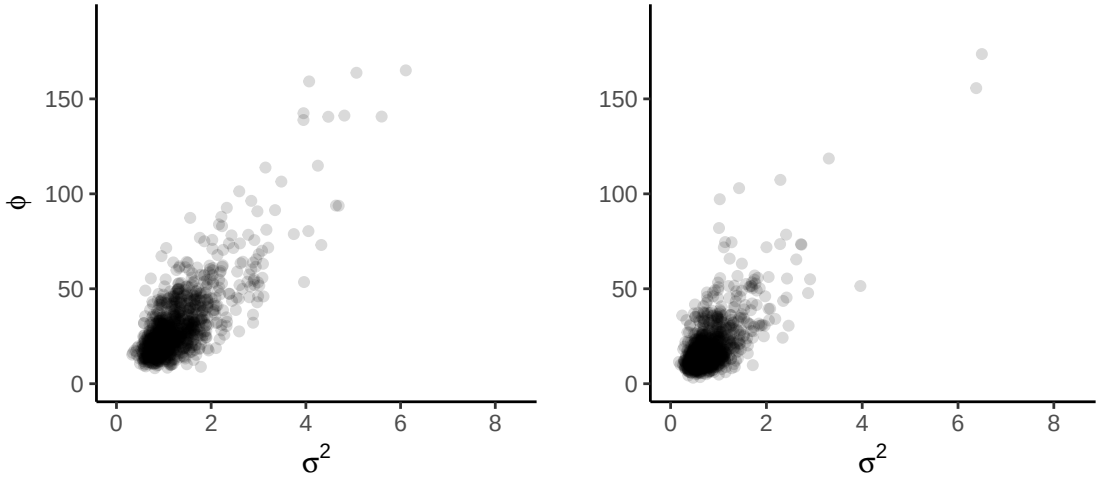


FIGURE 3.12: Posterior distributions of β for all Models



(A) Posterior distributions of σ^2 (left) and ϕ (right) for Model 3 and Model 4



(B) Posterior pairs between σ^2 and ϕ for Model 3 (left) and Model 4 (right)

FIGURE 3.13: Posterior distributions for σ^2 and ϕ

While comparing WAIC and LOO-CV, we can see that Model 3 and 4 are better than Model 1 and 2 which implies that imposing spatial structure is useful (see Table 3.8). And Model 3 is slightly better than Model 4.

Figure 3.14 shows the posterior fitted values against the observations. We can see that, imposing spatial structures in Model 3 and Model 4 help us to capture more information about the relations between locations than the non-spatial structure models. Also, the properties of gamma spatial model (see Section 2.3) reduce the uncertainty of when the value is small compared to non-spatial models.

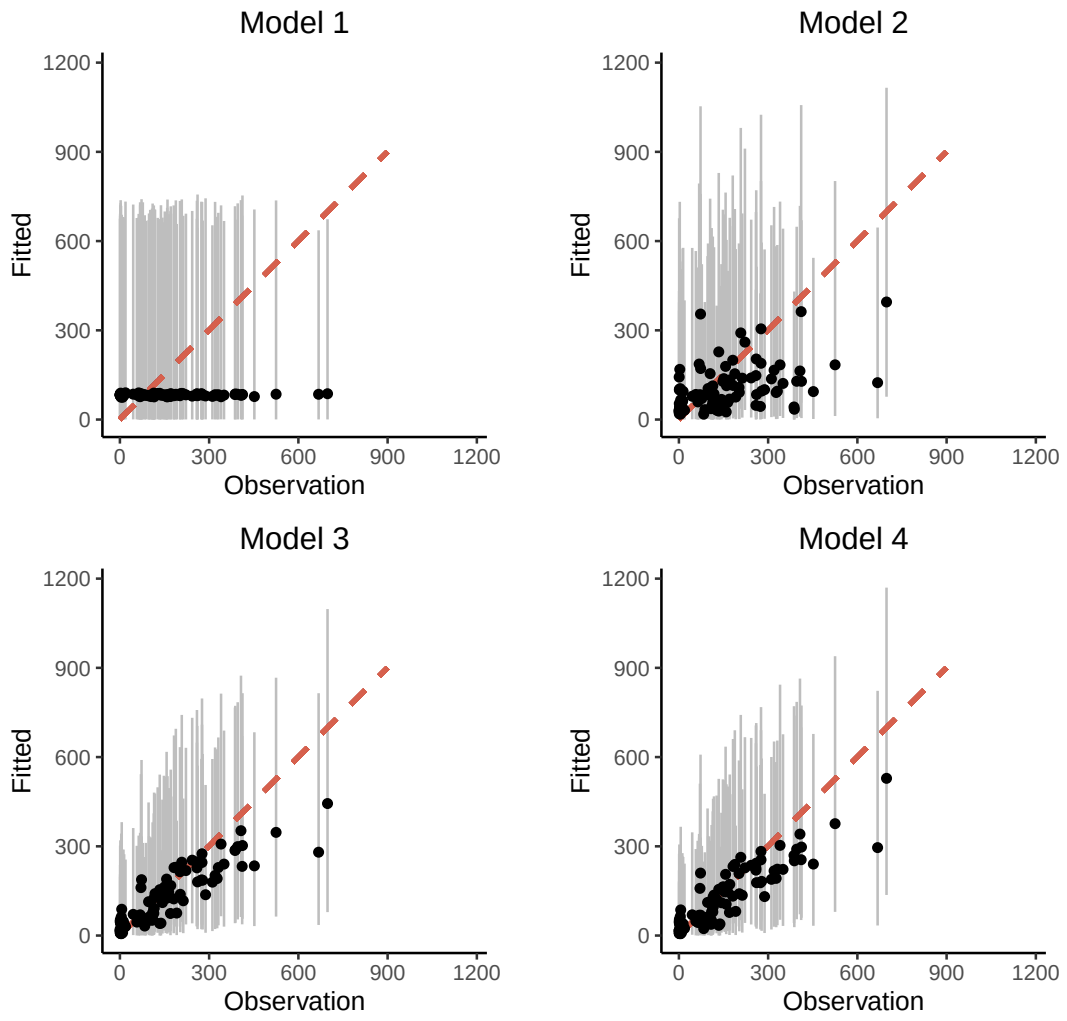


FIGURE 3.14: Fitted vs Observation values for Model 1 to Model 4. Red dashed line is fitted equals observations. "•" is posterior median of the fitted against observations, gray vertical line is 95% credible intervals.

We may also examine the residuals for all models. Figure 3.15 shows the residuals for predicted As level for all models at observed locations. We can see that in terms of median, none of these residuals appear obvious patterns. We may confirm that the gamma model is appropriate. Moreover, from the spread median residuals we can see that Model 3 and Model 4 more stable residuals than Model 1 and Model 2.

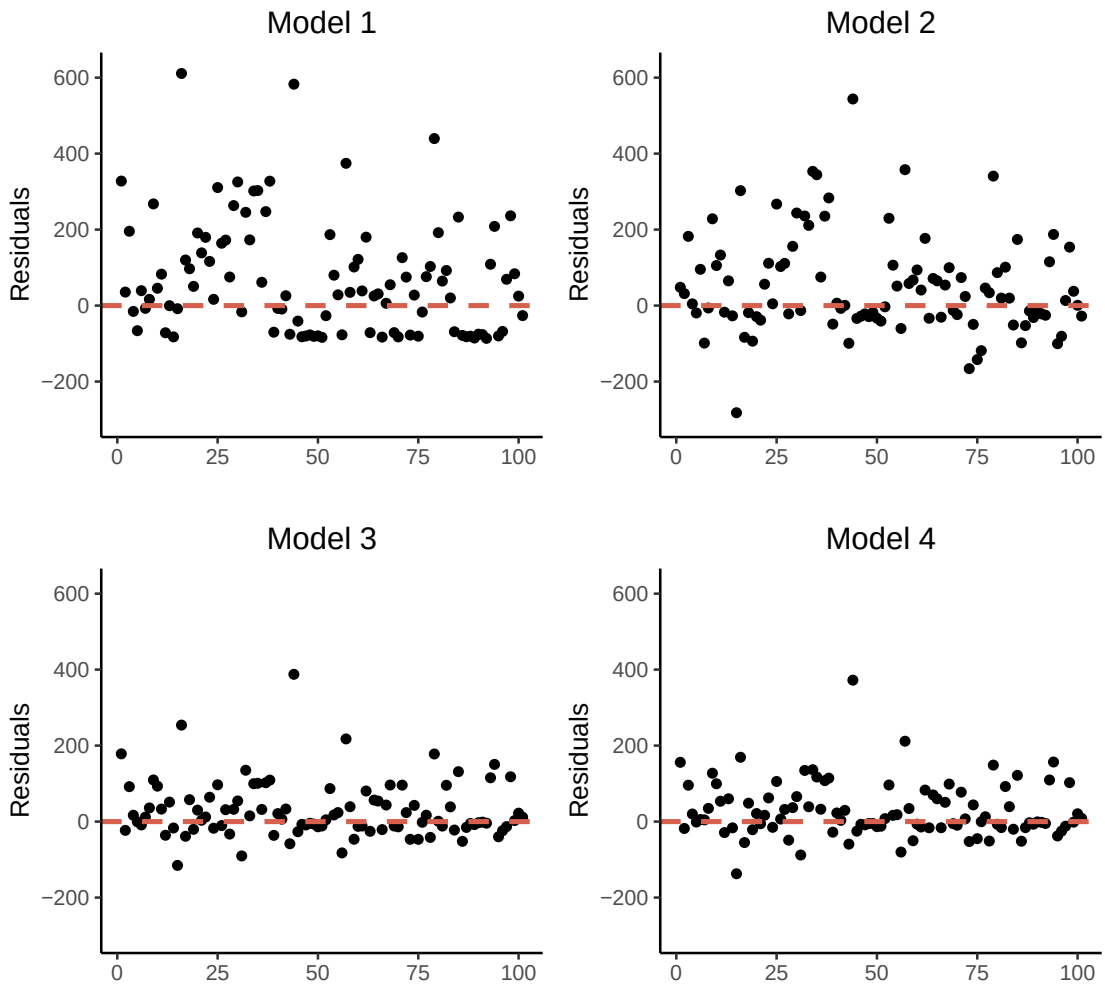


FIGURE 3.15: Residuals for Model 1 to Model 4. "•" is posterior median

3.2.3 Kriging

We perform kriging with the parameters obtained from Model 3 and Model 4. In the demonstration, we take the interval of the new location grid to be 1 km, the total number of new locations is $n_{\text{new}} = 3131$. The kriging for a subset of 1000 samples (recall total iterations is 2250) takes about 90 minutes. It is estimated that kriging may take more than 6 month to complete if we treat $\mathbf{y}_{\text{new}}$ as missing values.

We first illustrate the kriging on the spatial process, i.e. $Z(\mathbf{s}_{\text{new}})$. This follows Algorithm 4 to line 10. Figure 3.16 shows the median of the kriging surface for Model 3 and Model 4. We put both model in the same color scale and ignore the surface of standard deviation because it remains almost constant across the whole region. Clearly, because the model is accounting for spatial structure, we can see that there is a clear spatial structure for both models. But is also notable that the variation of color in Model 3 is more

obvious than that in Model 4. This may implies that Model 3 carries bigger variation across the region. This is logical, because Model 3 assume a constant mean structure while Model 4 introduces easting and northing as covariates. In Model 3, the all variation of the process has been carried by the spatial process. This also leads to the larger spread of the hyper-parameter pair $\{\sigma^2, \phi\}$. For Model 4, in contrast, beside the spatial structure, the covariates also carry information about spatial structure. Hence these covariates also help the model to decompose the source of variation, which resulted a smaller spread of the pair $\{\sigma^2, \phi\}$.

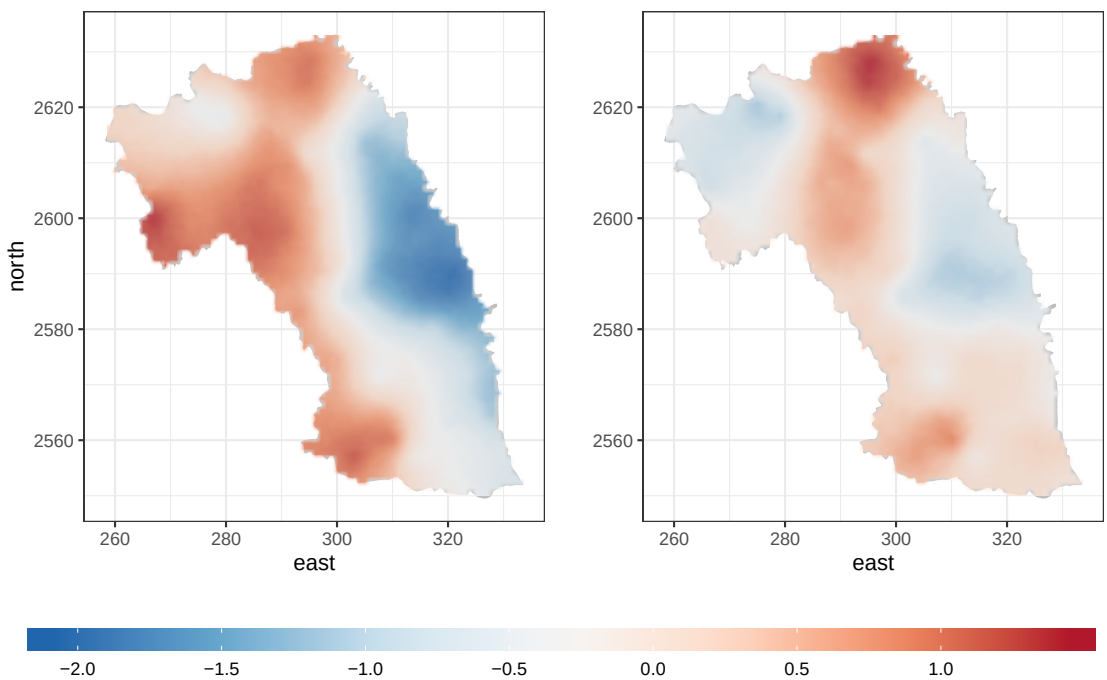
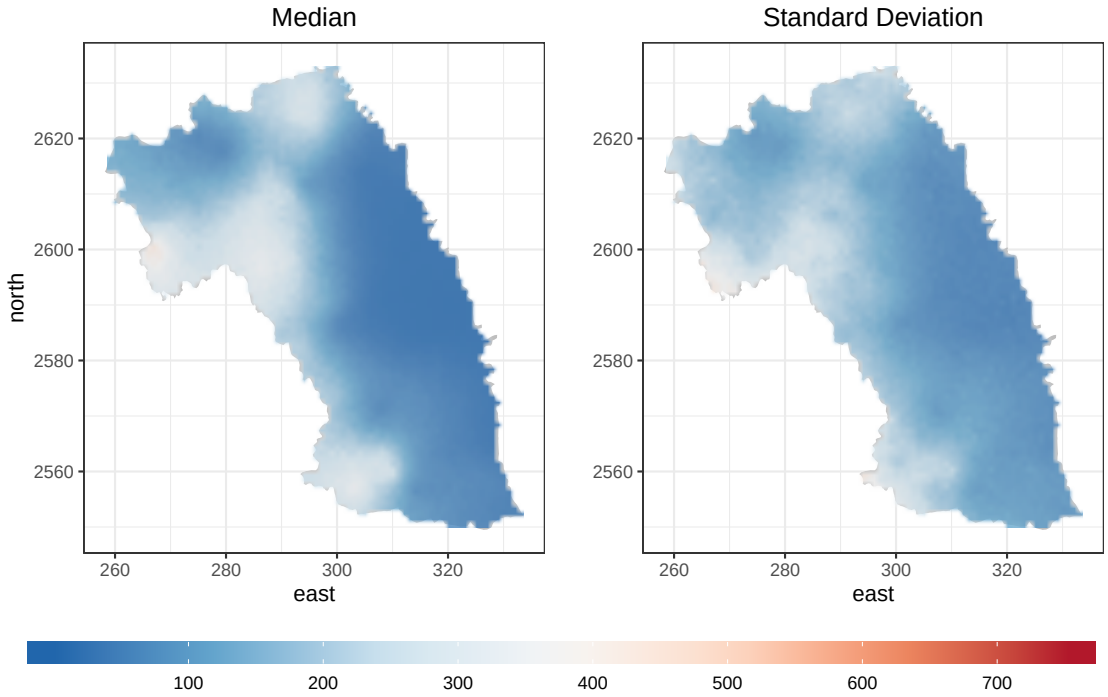
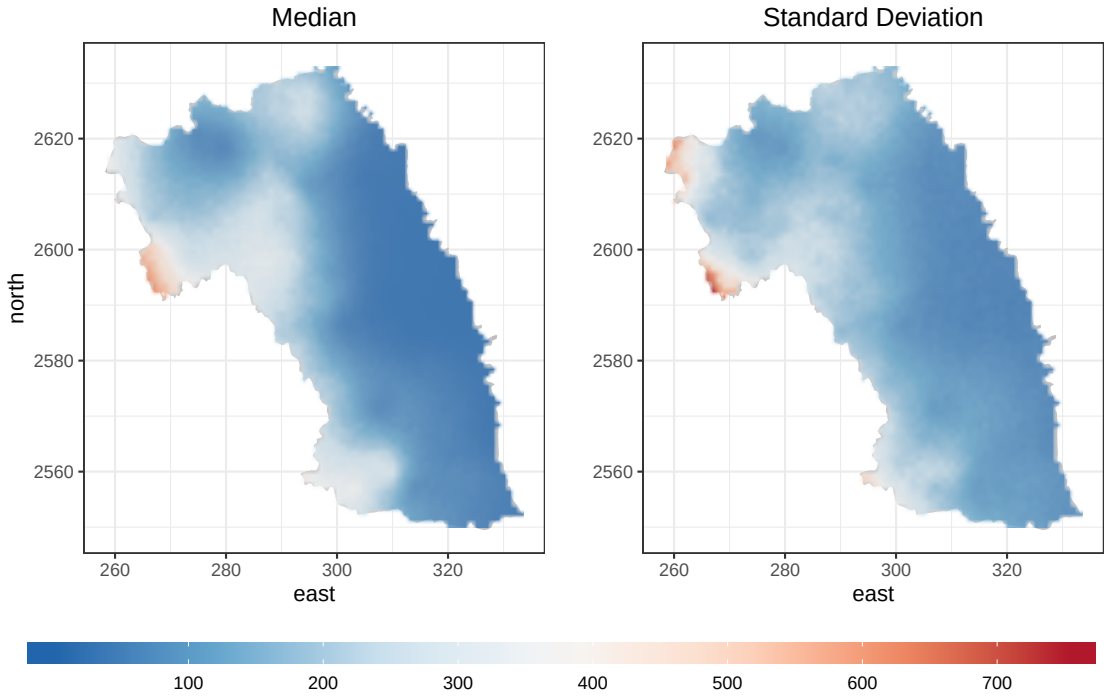


FIGURE 3.16: Median of kriging surface for spatial process for Model 3 (left) and Model 4 (right)

Figure 3.17 illustrates the kriging surface of Comilla As levels based on 1000 iterations. The left hand side is the median of the predicted levels, while right hand side is the standard deviation of the predictions. Similar to the kriging on the spatial process, Figure 3.17a and 3.17b share the same scale in color scheme. We can see that in terms of median, both model give similar results. But Model 4 has larger standard deviation than Model 3. This is also supported by WAIC and LOO-CV.



(A) Model 3



(B) Model 4

FIGURE 3.17: Kriging Surface for As level in Comilla

3.3 Comparison with other models

There are other distributions that handle skewness such as the Inverse Gaussian. Zhang and El-Shaarawi, 2010 proposed a skew-Gaussian processes model and Zareifard et al., 2018 proposed a model called Gaussian-log Gaussian convolution (GLGC).

Inverse Gaussian Inverse Gaussian is a two parameter distribution, with $\mu > 0$ being mean and $\lambda > 0$ indicating the shape. The density of $\mathcal{IG}(\mu, \lambda)$ is

$$p(y|\mu, \lambda) = \sqrt{\frac{\lambda}{2\pi y^3}} \cdot \exp\left\{-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right\}, y > 0 \tag{3.1}$$

The mean and variance of the Inverse Gaussian distribution are:

$$\mathbb{E}(y) = \mu \qquad \text{var}(y) = \frac{\mu^3}{\lambda}$$

For simplicity, we use log as link function (canonical link function is μ^{-2}). The hierarchical structure of the model is,

$$\begin{aligned} Y(\mathbf{s}_i) &\sim \mathcal{IG}(\mu(\mathbf{s}_i), \lambda) \\ \ln \alpha(\mathbf{s}_i) &= \mathbf{x}(\mathbf{s}_i)^\top \boldsymbol{\theta} + Z(\mathbf{s}_i) \\ Z(\mathbf{s}) &\sim \text{GP}(\mathbf{0}, \boldsymbol{\Sigma}) \end{aligned} \tag{3.2}$$

Skew-Gaussian processes The Skew-Gaussian process model relaxes the spatial structure of a Gaussian process (multivariate normal) into a skew-Gaussian process (multivariate skew-normal), allowing the spatial structure to capture the skewness. Let $y_1, y_2 \overset{\text{i.i.d}}{\sim} N(0,1)$, then for $\delta \in [-1, 1]$

$$z = \delta|y_1| + \sqrt{1-\delta^2} \cdot y_2 \tag{3.3}$$

is called a *skew-normal* distribution, where z is right skewed if $\delta > 0$. Note that $\delta^2 + (\sqrt{1-\delta^2})^2 = 1$. The multivariate extension of Equation (3.3) is given by Azzalini and Valle, 1996. For $\mathbf{y} = (y_1, \ldots, y_k)^\top$ with standardized marginals, independent of $y_0 \sim N(0,1)$ for $\delta_j \in [-1, 1], j = 1, \ldots, k$, define

$$z_j = \delta_j|y_0| + \sqrt{1-\delta_j^2} \cdot y_j \tag{3.4}$$

The joint density is called *multivariate skew-normal*. Then define

$$Z(\mathbf{s}) = \delta|Y_0(\mathbf{s})| + \sqrt{1-\delta^2} \cdot Y(\mathbf{s}) \tag{3.5}$$

Zhang and El-Shaarawi, 2010 remarked that despite the fact that each $Z(\mathbf{s})$ is skew-normal, but the "*finite dimensional distribution*" of $Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)$ is not multivariate skew-normal if $Y_0(\mathbf{s})$ varies with $\mathbf{s}$.

Recall that this is not a spatial GLM, hence we follow the structure of Equation (2.5). The skew- Gaussian model proposed by Zhang and El-Shaarawi, 2010 assumes that

$$Y(\mathbf{s}) = m_{\boldsymbol{\theta}}(\mathbf{s}) + \sigma_1 |X_1(\mathbf{s})| + \sigma_2 X_2(\mathbf{s}) + \sigma_0 \varepsilon(\mathbf{s}) \tag{3.6}$$

where $\sigma_0 > 0$ indicates the weight for nugget effect and $\sigma_2 > 0$. $\sigma_1 \in \mathbb{R}$ but in our case $\sigma_1 > 0$ since we are dealing with right-skewed data. Finally we provide the hierarchical model structure of Equation (3.6)

$$\begin{aligned} &Y(\mathbf{s}_i) \sim N(\mu(\mathbf{s}_i), \tau^2) \\ &\mu(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)^\top \boldsymbol{\theta} + \sigma_1 |Z_1(\mathbf{s}_i)| + \sigma_2 Z_2(\mathbf{s}_i) \\ &Z_r(\mathbf{s}) \sim \mathrm{GP}(\mathbf{0}, \boldsymbol{\Sigma}_r), r = 1, 2 \end{aligned} \tag{3.7}$$

note that $Z_1(\cdot)$ and $Z_2(\cdot)$ are independent, we may assign different correlation functions to $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. For simplicity, we may also assign same correlation function, say, exponential correlation function for $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ as for gamma model.

Gaussian-log Gaussian convolution Instead of using the absolute value of a Gaussian process to capture skewness, Zareifard et al., 2018 propose to use a log-Gaussian process, they argue that for the skew-Gaussian process,

1. Since the absolute function is not monotone, data usually do not carry enough information to recover the parameters related to $X_1(\mathbf{s})$ in Equation (3.6).

2. Whether the process $\mu(\mathbf{s})$ is mean square differentiable is not guaranteed by $X_1(\mathbf{s})$ and $X_2(\mathbf{s})$ in Equation (3.6).

Zareifard et al., 2018 propose not using the absolute function, but using exponential function instead to guarantee the monotonicity and differentiability of the function. Replacing $|X(\mathbf{s})|$ by $\exp(X(\mathbf{s}))$ and σ_r is the same as skew-Gaussian model. The hierarchical form of the GLGC model is

$$\begin{aligned} &Y(\mathbf{s}_i) \sim N(\mu(\mathbf{s}_i), \tau^2) \\ &\mu(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)^\top \boldsymbol{\theta} + \sigma_1 \exp\left[Z_1(\mathbf{s}_i)\right] + \sigma_2 Z_2(\mathbf{s}_i) \\ &Z_r(\mathbf{s}) \sim \mathrm{GP}(\mathbf{0}, \boldsymbol{\Sigma}_r), r = 1, 2 \end{aligned} \tag{3.8}$$

The setting of of $\boldsymbol{\Sigma}_r$ can be similar to the Equation (3.7).

We set up the hierarchical models following Equations (3.2), (3.7) and (3.8) and assigning similar priors to the corresponding parameters and then compare the posterior fitted samples against the observations of each models.

For simplicity, all models are based on observed As levels and the sample size is $n_{\text{obs}} = 101$. We all use same exponential correlation function as spatial structures and generate $\mathbf{y}_{\text{obs}}$, this is similar to what we have compared in Figure 3.14.

Figure 3.18 illustrates the results under different models. The gamma model on the top-left is Model 4 from our study. Note that the y axis of each plot is not in the same scale.

For Inverse Gaussian, we can see that the uncertainty (95% credible interval) of the fitted value is of higher order in scale compare to all three other models, the highest value reaching close to $50,000$. Moreover, when we incorporate left censoring, the MCMC seems unstable.

For the skew-Gaussian, the uncertainty is of similar scale to the gamma model. The posterior medians of the fitted values underestimate the true value when the value increase. Moreover, the uncertainty of $\mathbf{y}_{\text{obs}}$ covers negative values which violates the fact that all observations should be positive.

The GLGC model has similar situation, and the underestimation of the fitted values seems worse in the case of the skew-Gaussian model.

By comparing this to the result of the competitors, we can see that gamma model is able to capture some information of the skewness and the other competitors either have too much uncertainty or produced unreasonable predictions, although the uncertainty of the gamma model is seemingly large. However, after comparing other models and specifying different priors, we may say that, under this correlation structure, for now that is the best a gamma model can achieve.

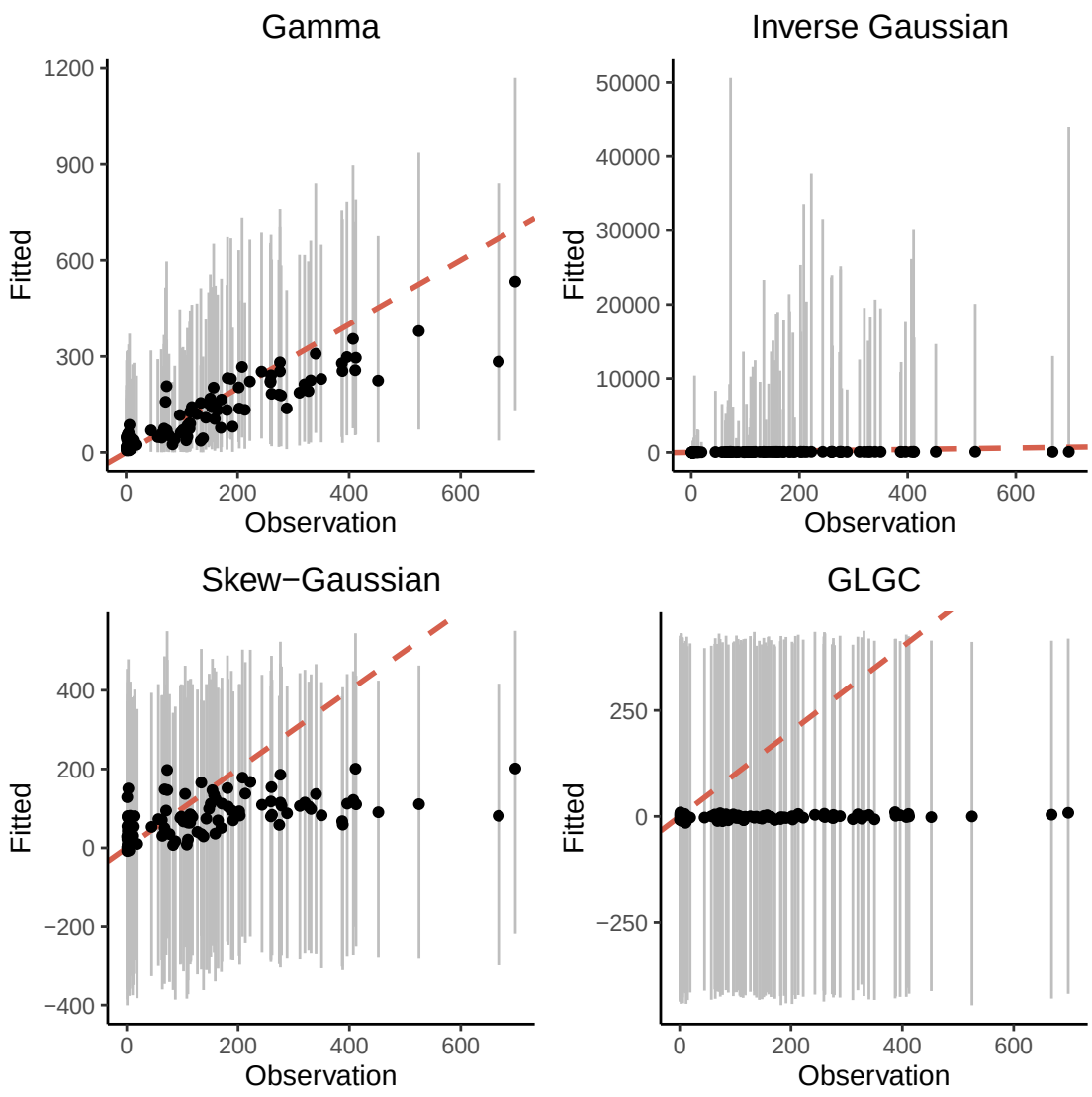


FIGURE 3.18: Fitted vs Observation values for Gamma, Inverse Gaussian, Skew-Gaussian and Gaussian-log Gaussian convolution. Red dashed line fitted equals observations. "•" is posterior median of the fitted against observations gray vertical line is 95% credible intervals.

Chapter 4

Conclusion and Future Work

4.1 Conclusion

We proposed a Gamma spatial model to account for two types of censored data to model the arsenic level in Comilla, Bangladesh. We carefully examined the prior specification for the scale parameters to choose a reasonable prior distributions. We then fit the models and compare different mean structures under log scale and then choose the best models to perform kriging over the whole area of Comilla. Finally, we compared our gamma models with other skew-distributed models in the literature.

Our conclusion is as follows: for the data in Comilla, gamma spatial model(s) is the best model to describe the skewness and provide reasonable predictions. The spatial structure is necessary in the gamma model(s), it helps us to capture the geographical relations. For the Gamma spatial models, assuming a constant mean structure (Model 3) gives a slightly better result than accounting easting and northing (Model 4) in terms of the variation, but the performance are close.

4.2 Limit

During the study, we realize that there are some restrictions limit our models.

In our models (Model 2 and Model 4), we use easting and northing as covariates, there are some difficulties applying other covariates.

Well depth We have the data for well depth for each location. we fitted the models and included well depth and it shown to be significant (95% credible do not include 0). However, we have to drop this useful information when performing kriging, because for new locations, we do not have the depth information. Predicting the depth will introducing new measurement errors, which may cause greater uncertainty for the predictions.

River It is reasonable to assume that the arsenic levels will be high when the location is close to the rivers. In our study, we haven't incorporated this information into covariates for the following two reasons: we were not able obtain the river information until the very end of this study. Moreover, there may be more than one river in Comilla. What are the appropriate way to incorporate the river information, e.g. defining one river as major river and ignore the sub-stream or assigning rivers in different weights according to other information such as flux, width, etc. Such incorporation may require further exploration.

Terrain Yu, Harvey, and Harvey, 2003 have considered the types of terrains. However, we could not find out the corresponding information on the BGS website and hence we were not able to incorporate these information in the data. We have the following reasoning for the district Comilla: if the type of terrain are the same across all Comilla, then because this covariates will served as a categorical variable to the mean structure, hence it is a constant, the final result will not have to much difference. Otherwise the information of different types of terrain may help us to decompose the source of variations.

4.3 Future Work

Correlation Funtion In this study, we only assume an exponential correlation function. There are also other possible correlation functions that are commonly used for spatial data.

Table 4.1 summarizes some common used correlation functions which will satisfy the covariance structure to be positive-definite. Notations styles follows Diggle, Tawn, and Moyeed, 1998

	Correlation function $\rho(d)$
Exponential	$\exp\{-d/\phi\}$
Power Exponential	$\exp\{-(d/\phi)^{\nu}\}$
Spherical	$\left[1-\frac{3}{2}(d/\phi)+\frac{1}{2}(d/\phi))^3\right], d \in [0,\phi]$
Wave	$(\phi/d)\cdot\sin(d/\phi)$
Matérn	$2^{1-\nu}/\Gamma(\nu)\cdot(d/\phi)^{\nu}K_{\nu}(d/\phi)$

TABLE 4.1: Summary of common isotropic correlation functions.

For the Matérn correlation function, $K_\nu(\cdot)$ denotes the is the modified Bessel function of the second kind. It has not yet been implemented in `Stan`. Using this correlation function may require coding the $K_\nu(\cdot)$ by ourselves.

New Models We may also explore new models that can handle skewed data. For data of the original scale, Xu and Genton, 2017 proposed a model through Tukey g-and-h distributions to construct a random field. On transformation of the data such as log scale or cubic root, we may also consider mixture of Gaussian models, the motivation comes from EDA, Figure 4.1 displays the histograms of Comilla data in log-scale and cubic root scale. For these models, however, managed to account for the censoring and how to choose reasonable priors for parameters are still challenging. Moreover, models for transformed data may loss the interpretation in terms of parameters indications.

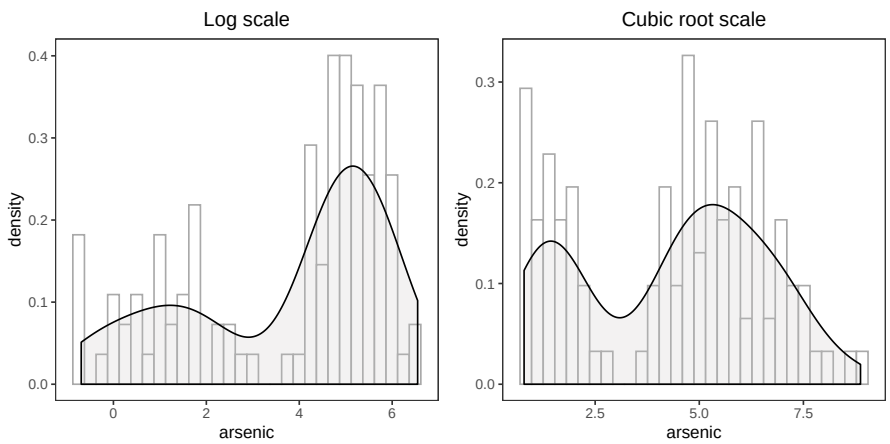


FIGURE 4.1: Transformation of the arsenic data in Comilla. LHS is on log-scale and RHS is on cubic root scale

Anisotropy One of the assumption that we made while fitting this models is the isotropy. However, this assumption in practice an be violated due to some geological and environmental factors. In our study, we believe by imposing the information of river flow (e.g. distance to rivers) may be more realistic. Some paper have been discussing relaxing the assumption of isotropy. For instance, Neto, Schmidt, and Guttorp, 2014 proposed an kernel function that makes use of the norm of projection measurements.

Large Scale Computation The data we have analyzed is a subset of the BGS survey. The complete data set contains a total number of $n = 3534$ observations. Computation through MCMC is very challenging for this scale of data.

The difficulty of spatial computation through Gaussian process is mainly due to inverting a high dimensional matrix. Datta et al., 2016 has proposed a class of highly scalable nearest-neighbor Gaussian process (NNGP) models to provide fully model-based

inference for large geostatistical datasets. He showed that the NNGP is a well-defined spatial process providing legitimate finite-dimensional Gaussian densities with sparse precision matrices then embed the NNGP as a sparsity-inducing prior within a rich hierarchical modeling framework and outline how computationally efficient MCMC can be executed without storing or decomposing large matrices. Zhang, Datta, and Banerjee, 2018 provided a demonstration of the NNGP for response and latent structure through a simulation study and they also illustrated a `Stan` implementation for the models[1].

We may adapt this technique into our gamma model or other spatial GLMs in the future to handle the computation of complete BGS data.

[1] https://mc-stan.org/users/documentation/case-studies/nngp.html

Appendix A

Implementations

A.1 `Stan` Implementation for Model 4

Name this model `m4.stan`, should be placed in the same folder as `R` script.

```
data {
int<lower=1> n_obs;
int<lower=1> n_cen1;
int<lower=1> n_cen2;
vector[n_obs] y_obs;
matrix[n_obs + n_cen1 + n_cen2, n_obs + n_cen1 + n_cen2] dist_matrix;

vector[n_obs + n_cen1 + n_cen2] mu_vec;
vector[n_obs + n_cen1 + n_cen2] east;
vector[n_obs + n_cen1 + n_cen2] north;
real L1;
real L2;
}

transformed data{
int N;
N = n_obs + n_cen1 + n_cen2;
}

parameters{
vector<lower=0, upper = L1>[n_cen1] y_cen1;
vector<lower=0, upper = L2>[n_cen2] y_cen2;
vector[3] b;
real<lower=0> beta;
real<lower=0> sigma;
real<lower=0> phi;
vector[N] noise;
}

transformed parameters{
vector[N] alpha;
real sigmasq;
```

```

sigmasq = sigma * sigma;

for (i in 1:N){
alpha[i] = exp(b[1] + b[2] * east[i] + b[3] * north[i] +  noise[i]);
 }
}

model {
// initialize covariance matrix and it cholesky decomposition
matrix[N,N] Sigma;
matrix[N,N] Lm;

// clean way of defining covariance matrix
Sigma = sigmasq * exp(-dist_matrix/phi);
Lm = cholesky_decompose(Sigma);

// assign priors, default will be flat priors
b ~ normal(0, 10);
beta ~ cauchy(0,2);

sigma ~ inv_gamma(2,1);
phi ~  inv_gamma(2,16);

// using cholesky deomposition to improve the speed
noise ~ multi_normal_cholesky(mu_vec, Lm);

// likelihood function
for (i in 1:n_obs){
y_obs[i] ~ gamma(beta*alpha[i], beta);
 }

// censored data
for (j in 1:n_cen1){
target += gamma_lcdf(L1| beta*alpha[n_obs + j], beta);
 }
for (j in 1:n_cen2){
target += gamma_lcdf(L2| beta*alpha[n_obs + n_cen1 + j], beta);
 }
}

generated quantities{
vector[N] y_fit;
vector[n_obs] log_lik;  // use for waic and loo-cv

// sample fitted values
for(i in 1:N){
y_fit[i] = gamma_rng(beta*alpha[i], beta);
```

```
 }

for(i in 1:n_obs){
// no density for censoring data, cannot record
log_lik[i] = gamma_lpdf(y_obs[i]| beta*alpha[i], beta);
 }
}

// need an empty line to end the program
```

Example code to run the model in `rstan`

```
library(rstan)
options(mc.cores = parallel::detectCores())
rstan_options(auto_write = TRUE)
Sys.setenv(LOCAL_CPPFLAGS = '-march=native')

data4 <- list(n_obs = n_obs, n_cen1 = n_cen1, n_cen2 = n_cen2,
y_obs = y_obs, dist_matrix = dist_matrix, mu_vec = mu_vec,
east = std_east_obs, north = std_north_obs, L1 = 6, L2 = .5)

m4 <- stan(file='m4.stan', data = data4, chains = 3,
iter = 1500, warmup = 750, refresh = 750)
```

A.2 R Implementation of Kriging: Algorithm 4

```
kriging <- function(stanfit, dist_all, east_new = NULL, north_new = NULL,
          num_of_sample = 1000){

# define names of extract parameters according to m4.stan
param <- c('b', 'beta', 'sigmasq', 'phi', 'noise')

num_of_param <- ncol(as.data.frame(rstan::extract(stanfit,
        pars = c('b', 'beta', 'sigmasq', 'phi'))))

# extracting parameters from stanfit object e.g. obtained from appendix A.1
p <- rstan::extract(stanfit, pars = param)
p <- data.frame(p$b, p$beta, p$sigmasq, p$phi, p$noise)

# justify which spatial model it is
if (num_of_param == 4){
colnames(p)[1:num_of_param] <- c('b', 'beta', 'sigmasq', 'phi')
} else {
colnames(p)[1:num_of_param] <- c('b0', 'b1', 'b2',
                'beta', 'sigmasq', 'phi')
}

# random choose a subset of parameters
total_samples <- nrow(p)
num_valid <- min(num_of_sample, total_samples)
row_idx <- sample(c(1:total_samples), size = num_valid, replace = FALSE)
p <- p[row_idx, ]

# get a complete matrix from old and new
num_obs <- stanfit@par_dims[["noise"]]
num_total <- nrow(dist_all)
num_new <- num_total - num_obs
end_idx_noise <- ncol(p)

D_obs <- dist_all[1:num_obs, 1:num_obs]
D_new <- dist_all[(num_obs + 1):num_total, (num_obs + 1):num_total]
D_A <- dist_all[1:num_obs, (num_obs + 1): num_total]
D_AT <- t(D_A)

# initialize matrix to store kriging values
y_new <- matrix(NA, nrow = num_valid, ncol = num_new)

east <- east_new
north <- north_new

# perform kriging using new location
for(i in 1: num_valid){
```

```
Sigma <- p[i, 'sigmasq'] * exp(-D_obs/p[i, 'phi'])

# fast way of computing symmetric matrix, Sigma_inv will be used twice
Sigma_inv <- chol2inv(chol(Sigma))

# parameters obtain from dataframe "p" defined earlier
noise <- as.numeric(p[i,(num_of_param + 1):end_idx_noise])

# calcluating conditional mean vector and covariance matrix
mu_new <- p[i, 'sigmasq'] * exp(-D_AT/p[i, 'phi'])  %*%  Sigma_inv %*% noise
Sigma_new <- p[i, 'sigmasq'] * exp(-D_new/p[i, 'phi']) - p[i, 'sigmasq']^2 *
    exp(-D_AT/p[i, 'phi'])  %*%  Sigma_inv %*% exp(-D_A/p[i, 'phi'])

# Rcpp (RcppArmadillo) library 10x faster computation than MASS::mvrnorm
noise_new <- mvnfast::rmvn(n = 1, mu_new, Sigma_new)
# relatively slow
# noise_new <- MASS::mvrnorm(n = 1, mu_new, Sigma_new)

if (num_of_param == 4){
alpha_new <- exp(p[i, 'b'] + noise_new)
} else {
alpha_new <- exp(p[i, 'b0'] + p[i, 'b1'] * east + p[i, 'b2'] * north + noise_
    new)
  }

# generate random gamma samples conditioning on spatial process
y_new[i,] <- rgamma(num_new, shape = p[i, 'beta'] *
          alpha_new, rate = p[i, 'beta'])
 }
return(y_new)
}
```

Appendix B

MCMC Diagnostic for Model 4

Inference for `Stan` model: m4. 3 chains, each with iter=1500; warmup=750; thin=1; post-warmup draws per chain=750, total post-warmup draws=2250.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	$\hat{R}$
θ_0	4.648	0.032	0.493	3.705	4.363	4.638	4.912	5.697	236.431	1.030
θ_1	-0.981	0.015	0.313	-1.582	-1.174	-0.982	-0.790	-0.354	447.249	1.003
θ_2	-0.449	0.015	0.315	-1.069	-0.650	-0.451	-0.244	0.171	427.018	1.000
β	0.013	0.000	0.003	0.008	0.011	0.013	0.015	0.021	486.003	1.005
σ^2	0.848	0.028	0.469	0.343	0.565	0.737	0.988	2.037	281.155	1.013
ϕ	20.399	0.818	14.894	6.520	11.543	16.301	24.206	60.335	331.374	1.006

TABLE B.1: Examples of `Stan` summary of parameters for Model 4 including diagnostic $\hat{R}$

Samples were drawn using NUTS (diag_e) at Sat Jul 13 22:30:40 2019. For each parameter, n_eff is a crude measure of effective sample size, and $\hat{R}$ is the potential scale reduction factor on split chains (at convergence, $\hat{R} = 1$).

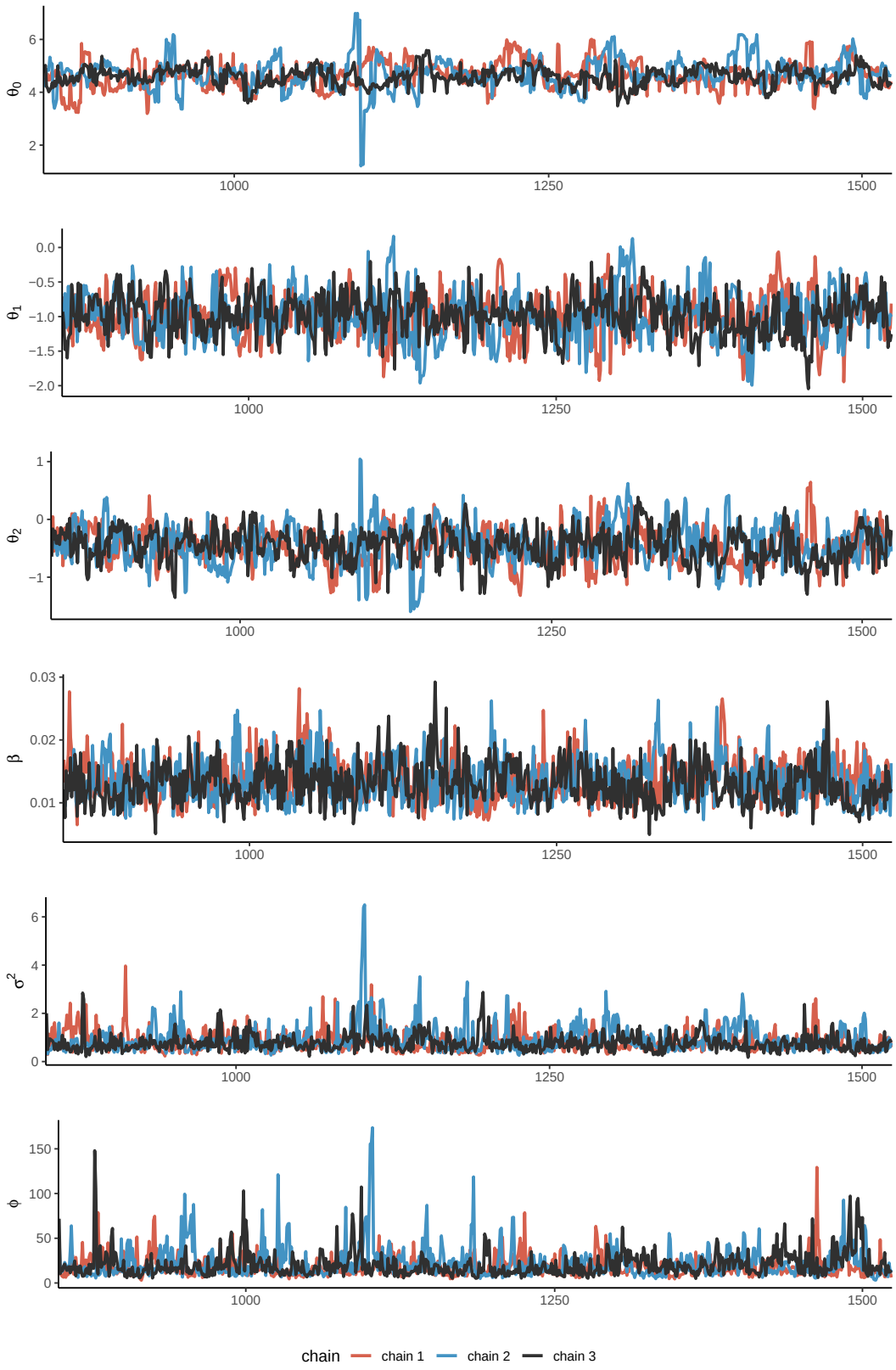


FIGURE B.1: Traceplot of Model 4, chains are pretty well mixed. Similar procedures are checked for other models.

Bibliography

Azzalini, Adelchi and A Dalla Valle (1996). "The multivariate skew-normal distribution". In: *Biometrika* 83.4, pp. 715–726.

Banerjee, Sudipto, Bradley P Carlin, and Alan E Gelfand (2014). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Betancourt, Michael (2017). "A conceptual introduction to Hamiltonian Monte Carlo". In: *arXiv preprint arXiv:1701.02434*.

Brooks, Steve et al. (2011). *Handbook of markov chain monte carlo*. CRC press.

Chakraborti, Dipankar et al. (2015). "Groundwater arsenic contamination in Bangladesh - 21 Years of research". In: *Journal of Trace Elements in Medicine and Biology* 31, pp. 237–248.

Cressie, Noel (1992). "Statistics for spatial data". In: *Terra Nova* 4.5, pp. 613–617.

Datta, Abhirup et al. (2016). "Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets". In: *Journal of the American Statistical Association* 111.514, pp. 800–812.

Diggle, Peter J, Jonathan A Tawn, and RA Moyeed (1998). "Model-based geostatistics". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47.3, pp. 299–350.

Flanagan, Sara V, Richard B Johnston, and Yan Zheng (2012). "Arsenic in tube well water in Bangladesh: health and economic impacts and implications for arsenic mitigation". In: *Bulletin of the World Health Organization* 90, pp. 839–846.

Gaus, I et al. (2003). "Geostatistical analysis of arsenic concentration in groundwater in Bangladesh using disjunctive kriging". In: *Environmental geology* 44.8, pp. 939–948.

Gelman, Andrew (2004). "Parameterization and Bayesian modeling". In: *Journal of the American Statistical Association* 99.466, pp. 537–545.

Gelman, Andrew, Jessica Hwang, and Aki Vehtari (2014). "Understanding predictive information criteria for Bayesian models". In: *Statistics and computing* 24.6, pp. 997–1016.

Gelman, Andrew, Donald B Rubin, et al. (1992). "Inference from iterative simulation using multiple sequences". In: *Statistical science* 7.4, pp. 457–472.

Gelman, Andrew et al. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)". In: *Bayesian analysis* 1.3, pp. 515–534.

Gelman, Andrew et al. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Hoffman, Matthew D and Andrew Gelman (2014). "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." In: *Journal of Machine Learning Research* 15.1, pp. 1593–1623.

Hossain, Mohammed Faruque (2006). "Arsenic contamination in Bangladesh - an overview". In: *Agriculture, ecosystems & environment* 113.1-4, pp. 1–16.

Kinniburgh, DG and PLeds Smedley (2001). "Arsenic contamination of groundwater in Bangladesh". In:

Kruschke, John (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Metropolis, Nicholas et al. (1953). "Equation of state calculations by fast computing machines". In: *The journal of chemical physics* 21.6, pp. 1087–1092.

Militino, Ana F and M Dolores Ugarte (1999). "Analyzing censored spatial data". In: *Mathematical Geology* 31.5, pp. 551–561.

Neto, Joaquim Henriques Vianna, Alexandra M Schmidt, and Peter Guttorp (2014). "Accounting for spatially varying directional effects in spatial covariance structures". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 63.1, pp. 103–122.

Polson, Nicholas G, James G Scott, et al. (2012). "On the half-Cauchy prior for a global scale parameter". In: *Bayesian Analysis* 7.4, pp. 887–902.

Rathbun, Stephen L (2006). "Spatial prediction with left-censored observations". In: *Journal of agricultural, biological, and environmental statistics* 11.3, p. 317.

Schmidt, Alexandra M., Kelly C. M. Gonçalves, and Patrícia L. Velozo (2017). "Spatiotemporal models for skewed processes". In: *Environmetrics* 28.6, e2411. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2411.

Sen, Pranab K (2016). "Abundant Environmental Arsenic Contamination: Some Statistical Perspectives". In: *Sankhya B* 78.2, pp. 341–361.

Snyder, John P (1987). "Map projections—a working manual (US geological survey professional paper 1395)". In: *United States Government Printing O ce, Washington, DC* 20402, p. 10.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry (2017). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". In: *Statistics and computing* 27.5, pp. 1413–1432.

Watanabe, Sumio (2010). "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory". In: *Journal of Machine Learning Research* 11.Dec, pp. 3571–3594.

Xu, Ganggang and Marc G Genton (2017). "Tukey g-and-h random fields". In: *Journal of the American Statistical Association* 112.519, pp. 1236–1249.

Yu, Winston H, Charles M Harvey, and Charles F Harvey (2003). "Arsenic in groundwater in Bangladesh: A geostatistical and epidemiological framework for evaluating health effects and potential remedies". In: *Water Resources Research* 39.6.

Zareifard, Hamid et al. (2018). "Modeling Skewed Spatial Data Using a Convolution of Gaussian and Log-Gaussian Processes". In: *Bayesian Analysis* 13.2, pp. 531–557.

Zhang, Hao and Abdel El-Shaarawi (2010). "On spatial skew-Gaussian processes and applications". In: *Environmetrics: The official journal of the International Environmetrics Society* 21.1, pp. 33–47.

Zhang, Lu, Abhirup Datta, and Sudipto Banerjee (2018). "Practical Bayesian modeling and inference for massive spatial datasets on modest computing environments". In: *arXiv preprint arXiv:1802.00495*.