UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXII

A Divide and Conquer Approach for Large Spatial Dataset

Coordinatore del Corso: Prof. Massimiliano Caporin

Supervisore: Prof. Carlo Gaetan, Cá Foscari University of Venice, Italy

Co-supervisore: Dr. Emanuele Giorgi, Lancaster University, UK

Dottorando/a: Md Moinuddin

02 December 2019

Abstract

In recent times, the rise of 'big data' has brought along major computational challenges in all the main disciplines of scientific research, including the field of spatial statistics. Some of these challenges include parametric estimation and quantification of estimation uncertainty that, when building statistical models using big data, pose an important computational load. Many methods have been proposed to address these challenges such as dimension reduction, approximation by Markov random fields, tapering of the covariance matrix, and subsampling based approaches. In this thesis a new *divide-and-conquer* approach is proposed that we call `farmer` for providing effect size and standard error estimates in spatial models of big data. According to the proposed approach, all observations are divided into blocks that are mutually exclusive according to their position. For each block, the model parameters are estimated and recombined using a fixed or random meta-model to take into account the (possible) spatial dependence. This generalized method can be applied to a wide range of spatial models. For example, consider a linear Gaussian spatial model. In a simulation study, the `farmer` estimators were compared with estimators based on methods with similar sampling ideas. In the context of the Gaussian model, two applications with real data are presented. The proposed method appears computationally efficient compared to equivalent methods and has lower bias in the estimates. Furthermore, the proposed approach provides a more realistic estimate of standard errors. Finally, we propose an application of the method to generalized linear spatial models for simulated and real counting data.

Sommario

Negli ultimi due decenni l'avvento dei *big-data* ha portato sfide computazionali in tutte le principali discipline della ricerca scientifica. Anche la Statistica spaziale sta affrontando questa sfida. Quando un modello parametrico viene proposto per *big-data*, la stima parametrica e la quantificazione dell'incertezza nella stima comporta un carico computazionale importante. Per questo sono stati proposti molti metodi per gestire queste sfide quali la riduzione della dimensionalità, l'approssimazione mediante campi casuali di Markov, la rastremazione *tapering* della matrice di covarianza e approcci basati sul campionamento. In questa tesi si propone un nuovo approccio *divide-and-conquer* detto `farmer` per la stima e la valutazione dell'incertezza dei parametri in modelli spaziali in presenza di grandi moli di dati spaziali. Secondo l'approccio proposto tutte le osservazioni vengono divise in blocchi mutualmente esclusivi secondo la loro posizione e per ogni blocco si stimano i parametri del modello. Le stime vengono quindi ricombinate tramite un meta-modello a effetti fissi o casuali per tenere conto della (eventuale) dipendenza spaziale. Il metodo risulta completamente generale e pu essere applicato ad un ampia gamma di modelli spaziali A titolo d'esempio viene considerato un modello spaziale lineare gaussiano. In uno studio di simulazione gli stimatori `farmer` sono stati confrontati con stimatori che si basano sulla medesima idea di campionamento Sempre nel contesto del modello gaussiano si presentano due applicazioni con dati reali. Il metodo proposto è risultato computazionalmente efficiente rispetto ai metodi concorrenti, con distorsione delle stime inferiore. Inoltre, l'approccio proposto fornisce una stima più realistica degli errori standard. Infine si propone un'applicazione del metodo a modelli spaziali lineari generalizzati per dati di conteggio simulati e reali.

Dedication

to Shahnaz, Thubaita & Omar

Acknowledgements

I am extremely grateful to have had the opportunity to work with Prof. Carlo Gaetan. The approach to statistical research he has taught me is eminently practical, principled, and effective.

I would like to express my gratitude to Dr. Emaneuele Giorgi for his guidance, teaching how to identify and solve problems efficiently, warm and cordial supervision. I am also grateful to CHICAS for providing me an excellent working environment and facilities.

I would like to thank the reviewers for their positive appraisals, suggestions, and intellectual criticisms. These helped me a lot to improve the piece of work up to the mark.

I am grateful to my respected parents, my life partner, my daughter Thubaita and son Omar for their sacrifices. They have provided me with moral and emotional support in my life. I am also grateful to my other family members and friends who have supported me along the way.

I would like to thank my fellow doctoral students for their feedback, cooperation and of course friendship.

I am also grateful to Patrizia Piacentini, Ph.D. secretary for her unfailing support and assistance during the study period.

And finally, last but by no means least, also to everyone in the Statistics Department, it was a great sharing department with all of you during the last three years.

Thanks for all your encouragement!

Contents

List of Figures

List of Tables

Introduction

Overview

During the last few decades, there has been a data explosion by virtue of technological advancement. Mayer-Schönberger and Cukier (2013) explained nicely how every simple digital device generates millions of data points. Some of the fields where the *datafication* has made significant impacts include neuroscience, astronomy, nanoscience, finance and business, transportation, biology and medicine, health-care and environments.

More than three billion search queries are received by the Internet giant Google every day. Google engineers make use these data in innovative ways; for e.g. they have shown how the search query data can be used in predicting the spread of flu in the United States even at the regional level (Ginsberg *et al.* (2009)). Another giant data generator is Facebook, that receives more than 10 million new photos with more than three billion comments and clicks in like button Mayer-Schönberger and Cukier (2013). Uber - a digital tech start-up that uses a geographic information system to make everyday transport easy for millions of people around the world, has operations in 785 metropolitan areas worldwide, and manages millions of geographic locations every day. Uber's engineers introduced `kepler.gl`, which is an open source large scale geospatial data-agnostic, high-performance web-based toolbox to get actionable insights from beautiful maps (He (2018)). An additional big data source is the large online marketplaces such as Amazon, AliExpress, and the air agencies that receive millions of purchase orders every day globally. These start-up techs have leveraged big data analytics to reshape their business policies for profit maximization. Satellite images are a rich source of information for astronomers; they

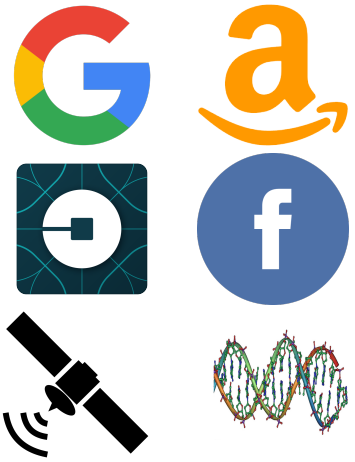

FIGURE 1: Some big data sources

have now also become a vast trove of data for other uses. For instance, from satellite images many environmental data such as carbon dioxide (CO_2), carbon monoxide (CO) and total column ozone (TCO) are being extracted. Among the myriad data types, genetics, brain imaging technologies, and other multi-dimensional databases are being generated at an accelerated rate for analysis. The 4 Vs that characterize big data - volume, veracity, variety and velocity of data, are increasing rapidly and new companies are leveraging big data in the hopes of creating artificial intelligence engines to reliably answer the most important questions. Big data analytics has emerged as a billion dollar industry and the number of big data companies is growing faster than ever (Grover *et al.* (2018)). Fortunately or unfortunately all the aforementioned data sources are associated with geographic location. There is of course information in the geo-location however to be most useful, that information needs to be turned into data. To *datafy* the locations of nature, objects and people specialized techniques are needed. The location associated data are called *spatial* data and special branch of statistics has been born to handle this type of data. The first formal and classic text on this issue came out two decades back (Cressie (1992)).

The three presidents of the American Statistical Association (ASA) mentioned big data as the big topic in the President's Corner of the June 2013 issue of AMStat News. The media coverage, conference announcements, celebration of the Big Data Week, and special initiatives taken from the White House drew attention of the world to the issue of big data. The presidents of ASA also motivated statisticians to come forward and collaborate more with other researchers for taking the lead in dealing with big data. This was also echoed by the president of Institute of Mathematical Statistics Yu (2014).

Kettenring (1997) has described six approaches to handle massive data. These include adaptive sampling, guided visualization, reliance on approximations, distributed work, divide and conquer, and exploit the context. During the last couple of decades, large number of methods focusing on these approaches have been introduced. Some of them are applicable to non-spatial settings only but some focused on large spatial data. The methods mostly aim to reduce computational cost. However some were designed for better prediction and statistical inference. The ideal methods would reduce computational cost, provide bias-reduced estimate, and realistic measures of standard error for the estimates when analyzing large spatial data types. In this thesis, we propose such an approach for handling large spatial data. The main contributions are described in the next section.

Main contributions of the thesis

In the current thesis we propose a new divide and conquer approach which we call `farmer` approach where er employ meta analysis techniques. Our proposed method is expected to be free from major shortcomings of the previous methods. **In the first contribution,** we have developed a divide and conquer approach for Gaussian random fields. According to our proposal we split the entire dataset into K mutually exclusive and exhaustive blocks. At each block we estimate the model parameters and respective covariance matrix using some method, such as maximum likelihood procedure. The point estimate $\hat{\theta}_i$ and respective covariance matrix $V(\hat{\theta}_i)$ at block i are then considered as the outcome from a i^{th} single study in meta analysis setting. In this way, we have K studies those are not necessarily independent. We propose to combine outcome of these studies using fixed and random effect meta analysis models(Hedges and Vevea (1998)). To estimate the parameters of the random effect meta analysis model we fit the multivariate conditional auto-regressive (MCAR) model. Fitting MCAR model is motivated by the fact that splitting the data into mutually exclusive and exhaustive blocks leaves us lattice. This approach provides analytical standard error of the global estimators. The spatial dependence parameters were negatively biased and we have proposed to apply bias correction technique proposed by Kosmidis *et al.* (2017) at block level which removes the bias. We have experienced empirical consistency of the estimators from simulation experiments. Also, we have compared the results with two existing methods and observed that `farmer` approach outperforms. We have applied the Gaussian geostatistical model to US precipitation data through `farmer` approach and found comparable results with MLE obtained from entire dataset. As a second application we applied trans-Gaussian model to river-blindness data over 18 African countries using proposed approach.

The proposed approach is intuitive, easy to apply, computationally efficient, feasible to parallelize, and provide bias reduced estimate and more realistic standard error of the estimates.

The second contribution in the thesis is to extend the approach for non-Gaussian large spatial data. For, non-Gaussian data there is no closed form nice likelihood function as in Gaussian case. This makes it more challenging to handle the large data. We proposed to employ the generalized linear geostatistical modeling approach (Diggle *et al.* (1998)) at block level and estimate the parameter using some methods such as, Monte Carlo maximum likelihood, Laplace approximation, hierarchical likelihood method or

generalized estimating equation. These methods allow us to find the point estimate $\hat{\theta}_i$ and covariance matrix $V(\hat{\theta}_i)$. We then proposed to apply the meta-analysis models as described before for obtaining the global estimates. For non-normal case since we do not have analytical form of the variance we cannot apply the bias correction method. We have shown application of binomial logistic model to river-blindness data again and obtained comparable results with MLE.

The rest of the thesis is organized in the following manner. In chapter 1, we have presented spatial big data computation methods and the room for improvements. The general formulation of the divide and conquer approach is also discussed in the same chapter. Chapter 2 will describe the mathematical details, implementation and other necessary adjustments of `farmer` approach. In the third chapter the performance of the `farmer` approach is evaluated using Monte Carlo simulation experiments and applied to two real life data sets. In chapter 4, the extension of the proposed approach is discussed for non-Gaussian data. Finally, in the last chapter, some concluding remarks are pointed out and the way forward are discussed.

Chapter 1

Spatial big data methods

1.1 Spatial big data methods

As mentioned, big spatial data is frequent in many fields. This chapter aims to provide a snapshot of methods available for spatial big data computation and explore the gap that needs to be addressed.

Before describing the methods we introduce first the notations used for spatial data using a Gaussian model. Let us assume $Y(s)$ is measurement of a random process at location $s; s \in \mathcal{D} \subset \mathbb{R}^2$. We are interested to study the behavior of the data generating process through $Y(s)$ measured at n different locations. The model of the process is defined as,

$$Y(s) = \mu(s) + S(s) + e(s) \tag{1.1}$$

where $\mu(s) = E(Y(s))$, and $S(s)$ is a stationary Gaussian spatial process which cannot be observed directly, also called *latent process*, with zero mean and covariance function $\sigma^2\rho(h_{ij}, \phi)$, where σ^2 is the variance parameter and ϕ is the scale or range parameter, h_{ij} is the distance between two points s_i and s_j. We assume that $e(s)$ is Gaussian white noise with variance τ^2, which is independent from $S(s)$. The τ^2 is also called *nugget* effect. Here, the process $S(s)$ represent the spatial variation in the process, on the other hand $e(s)$ explains the unstructured variation over and above sampling variation. The log likelihood function of the model is,

$$l(\theta, y) = -\frac{1}{2}n\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(y-\mu)^\top\Sigma^{-1}(y-\mu), \tag{1.2}$$

where $y = (y(s_1), \ldots, y(s_n))^\top$ is the vector of realization of $Y = (Y(s_1), \ldots, Y(s_n))^\top$, $\theta = (\mu, \sigma^2, \phi, \tau^2)$ is the vector of unknown parameters, $\mu = (\mu(s_1), \ldots, \mu(s_n))^\top$ is the

vector of mean, $\Sigma = \sigma^2 R(\phi) + \tau^2 I$ is the variance covariance matrix of Y where $R(\cdot)$ has ij^{th} element $R_{i,j} = \rho(h_{ij}, \phi)$, h_{ij} being the distance between i^{th} and j^{th} locations. Among many, we present two classes of correlation models here for convenience of the readers. The correlation models that we present here are Matérn class and Powered exponential class.

Matérn class

The Matérn class is a two parameter flexible class of correlation model which is defined as,

$$\rho(h) = \frac{1}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{h}{\phi}\right)^{\kappa} \mathcal{K}_{\kappa} \left(\frac{h}{\phi}\right),$$

where κ is the smoothness parameter, $\mathcal{K}_{\kappa}(\cdot)$ is the modified Bessel function of order κ, $\Gamma(\cdot)$ is the gamma function. The exponential correlation model is the special case of Matérn model when $\kappa = 1/2$. Also, the Gaussian correlation model is the limiting case of Matérn model, that is, when $\kappa \longrightarrow \infty$, Matérn $\longrightarrow$ Gaussian model. The graphical representation of the model is,

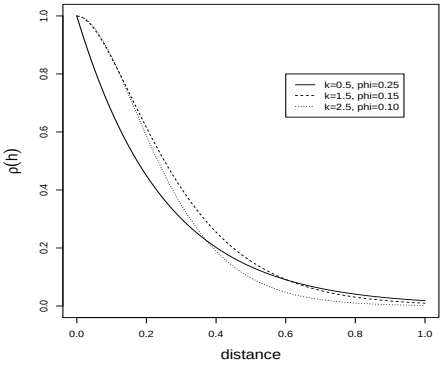

FIGURE 1.1: Matérn correlation model for three different values of scale and smoothness parameters.

Powered exponential class

The powered exponential class is also a two parameters family of correlation model which is defined as,

$$\rho(h) = \exp\left(-\frac{h}{\phi}\right)^{\kappa},$$

where κ is the smoothness parameter and ϕ is the scale. The exponential and Gaussian correlation models are the special case of powered exponential model when $\kappa = 1$ and

$\kappa = 2$ respectively. The graphical representation of the powered exponential model is shown below.

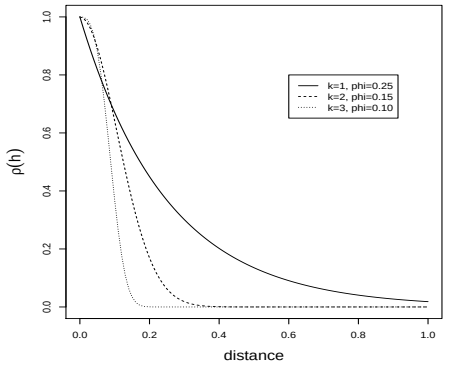

FIGURE 1.2: Powered exponential correlation model for three different values of scale and smoothness parameters.

Maximum likelihood estimation and assessing the quality of the estimates are computationally intractable for medium to large size spatial data. This is because evaluation of log likelihood function (1.2) requires $O(n^3)$ operations and $O(n^2)$ memory. The volume, veracity, variety and velocity of spatial data are increasing. These are the known four V's of big data. In case of spatial data there is another feature which is the reflection of Tobler's law. The law states that all the objects are associated but nearer objects are more alike than the distant one. This feature added an extra challenge to the statistician. Tobler's law enters into the analysis through $\rho(\cdot)$.

Many methods have been proposed to overcome the challenges of big spatial data. Broadly, they are based on covariance tapering, lower dimensional approximation, Markov random field approximation, composite likelihood-based approaches, and sub-sampling.

Tapering sets of covariance to zero deliberately after certain range which produce sparse linear system to solve in kriging setting(see for details Furrer *et al.* (2006), Stein (2013), Hirano and Yajima (2013)). Solving the sparse linear system is then become efficient to solve. The tapered covariance is defined as,

$$\Sigma_{tap} = \Sigma \circ \Sigma_{\delta},$$

where $\circ$ represents the direct or Schur product, Σ_{δ} is the sparse covariance matrix whose elements are zero after range defined by δ. Inverting Σ_{tap} instead of Σ is computationally efficient. However, all the covariance parameters can not be estimated consistently in this method. Also, tapering may not be effective when there exists long range correlation.

There are several developments focused on lower dimensional space approximation that utilize either Kalman filter, basis function or kernel convolutions (see for example Wikle and Cressie (1999), Banerjee *et al.* (2008), Cressie and Johannesson (2008)). In these methods the spatial process is approximated through lower dimensional space process. For example, in fixed ranked kriging, Cressie and Johannesson (2008) the spatial process are decomposed into a linear combination of K basis functions, such as

$$S(s) = \sum_{k=1}^{K} h_k(s)\theta_k,$$

where $h_k(s)$ is the basis function and θ_k is the coefficient of that function. By doing this we have to inverse $K \times K$ matrix instead of $n \times n$ where $K << n$. On the other hand, Stein (1999) and Fuentes (2007) have proposed to approximate the likelihood avoiding the matrix computation but rather working in the spectral domain of the process. For the approximation methods adequacy is always a concern. Approximating random fields by a Markov random field is another approach for approximation (see for example RUE and Tjelmeland (2002) Rue and Held (2005)). This approach uses the Markov property that is the distribution at a particular location and only depends on the observations at neighbor locations, simplifying the problem. While Markov random field is suitable for regular points over grid, it can also be fitted to irregular points by some modifications but at the cost of introducing some unquantifiable errors in precision.

Several statisticians have proposed to optimize the composite likelihood function instead of likelihood function (Curriero and Lele (1999), Bevilacqua *et al.* (2012), Eidsvik *et al.* (2014)). Composite likelihood is a general class of pseudo-likelihoods constructed based on marginal or conditional likelihoods of subsets of data. In this method, a set of marginal or conditional events are taken and the log-likelihood function for these observations set is constructed. The log-composite likelihood is then obtained by adding the individual log-likelihood from each set as if the components are independent. Curriero and Lele (1999) fist proposed to estimate the semivariogram parameters based on composite likelihood method. Later, Bevilacqua *et al.* (2012) proposed to use the weighted composite likelihood approach for estimating space and space-time covariance function. Instead of simply adding the component likelihoods, Bevilacqua *et al.* (2012) suggested weighted summation of the components. This approach allows us to strike a balance between computational complexity and statistical efficiency. Block composite likelihood approach is another proposal in this domain by Eidsvik *et al.* (2014). According to this approach entire domain is split into many smaller blocks and likelihood constructed in each block is considered as a component in composite likelihood setting. This approach

allows parallel computation. Sub-sampling is another approach for handling large data which is also known as the divide and conquer approach. Several studies have investigated this for non-spatial settings however few have examined in spatial settings. This method aims to split the entire data into many smaller subsets and estimate the model in the subset. The estimates obtained from subset are combined to obtain an overall estimate (see for example Liang *et al.* (2013), Barbian and Assunção (2017), Bickel *et al.* (2012), Chang *et al.* (2017), Zhang *et al.* (2015), Zhou and Song (2017)).

In the spatial setting, the few methods that exist include resampling based stochastic approximation (RSA) (Liang *et al.* (2013)), spatial subsemble (SpSub) estimator (Barbian and Assunção (2017)) and Meta-Kriging in Bayesian setting (Guhaniyogi and Banerjee (2018)). In the RSA method, a random subsample is drawn and the parameters set θ of the spatial model are estimated minimizing the Kullback-Leibler divergence and denoted as $\hat{\theta}^{(1)}$. A second subsample is drawn randomly and the estimate $\hat{\theta}^{(1)}$ is updated based on the second subsample using a set of equations. The set of equations are derived based on stochastic approximation of Kullback-Leibler divergence. The estimate obtained after updating in second step is denoted by $\hat{\theta}^{(2)}$. This process continues until convergence of $\hat{\theta}^{(k)}$ obtained, where $\hat{\theta}^{(k)}$ is the estimate from the k^{th} subsample. The authors have showed that under infill asymptotic, the final estimator $\hat{\theta}^{(k)}$ converges to the $\tilde{\theta}$, where $\tilde{\theta}$ is minimizer of Kullback-Leibler divergence for entire data set. The estimator have asymptotic normality as well; however there are some concerns about this method. Firstly, the subsample is taken randomly in the spatial setting shape and size of the subsample can affect the inference (see Lahiri (2013), Hall *et al.* (1995)). This important fact is completely ignored when the sample is selected. Secondly, two sequential subsample could have overlapped observations. In that case there should be an extra correlation between the subsample which is also ignored. Thirdly, the approach selects the samples sequentially and updates the previous estimates which does not allow implementation of the the process in parallel. Moreover, there is no clear guide to estimate the standard error (SE) of the estimates. Indeed, the process needs to be repeated many times to calculate the SE empirically, adding to the computational load. The SpSub method, on the other hand, select the spatially structured subsample randomly. At first, j centers are selected and around each centers, k nearest neighbors are selected therefore the subsample consists of jk observations. In this way the subsample contains nearer as well as distant observation. At the same time there is another center and same number of neighbor observations. This last small subsample are divided into two subsets, the validation subset $(Y(s)_v)$ and prediction subset $(Y(s)_p)$. The model parameter set θ are then estimated based on the main subsample using some method

such as maximum likelihood and the estimate is denoted by $\hat{\theta}_i$. The prediction is done for the validation subset $(Y(s)_v)$ based on $\hat{\theta}_i$ and $(Y(s)_p)$. The process is repeated B times and the global estimate is obtained by the weighted average, where the weight is the inverse prediction error at respective repetition. The standard error of the estimate are the weighted average of inverse Fisher information matrix calculated from subsets. The weight is the square of inverse prediction error. In this method, the prediction error is used as an weight for combining. However, the quantity of prediction error depends on the quality of the estimate $\hat{\theta}_i$ and the prediction subset data $(Y(s)_p)$. If the prediction subset is very similar to that of validation subset then prediction error could be lower even the quality of estimator is bad and vice versa. Another issue could be due to repeated observation in the two consecutive subsamples as described before. Also, there is no guarantee about the convergence of the estimators. Moreover, the scalability claimed by these methods assumes block-independence at some level but when the blocks borrow information across sub-regions, the scalability is lost. Furthermore, partitioning is always an issue.

The Meta-Kriging also splits the data and conduct Bayesian modeling at subset level to construct the posterior. The local posterior is then combined for obtaining the global posterior using geometric median approach. This method reduces computational cost but comes at a cost of less reliable inference. Also, the method has chance to miss the local feature of the spatial process and the between block dependence has been ignored.

In summary, although there are pitfalls, developments are ongoing in the large spatial data methods especially for estimating spatial dependence parameters. Some methods suffer from inconsistency and some from inadequacy and some are biased. In subsampling based methods, the between subsamples correlation has been ignored that possibly resulted in spuriously reduced standard errors. Moreover, non-Gaussian large data have not been dealt yet.

The current thesis aims to develop a new *divide and conquer* approach for big spatial dataset which consider the existing issues in the subsampling techniques. This approach aims to reduce the computational cost and provide more realistic standard error of the estimate. In this approach we propose to employ the fixed and random effect meta analytic tools for obtaining the global estimates suitably combining the local estimates. We named the approach as `"farmer"` approach due to its similarity to the technique that farmers in the developing countries adopt in harvesting large fields with limited resources. Also, since we apply the fixed and random effect meta analytic tools for obtaining the global estimators, the name `"farmer"` can be elaborated as **f**ixed **a**nd **r**andom effect **m**eta, **e**stimato**r** → **farmer**. Moreover, the author was a farmer during

1998 – 2005. The followings are two images from Bangladesh that resemble the existence of divide and conquer approach in nature.

FIGURE 1.3: (left) Paddy harvesting from large field with limited resource in Bangladesh, (right) natural splitting of large field.

In the left it is shown that how farmers harvest from their huge fields in the absence of big machines but employed many human in the same field. This is the real parallel work. At the right the splitting of large area based on ownership or due to feasibility of cultivation of land are shown. These motivated the name `farmer` approach.

In the next section we discuss very briefly the general formulation of divide and conquer approach.

1.2 Divide and conquer approach: general framework

Divide and conquer method is basically a two steps approach similar to `MapReduce` program (Yang *et al.* (2007)). In the first step, the large data set is split into smaller sets and in the second step the information obtained from each of the smaller sets are combined together. These information from smaller sets are obtained in the form of point estimates, uncertainty of the estimate or some function of the both. The targeted statistical procedures are employed over the smaller data sets for obtaining these information. This process of obtaining information could be titled as the *intermediate* step. The hypothetical splitting and generated lattice after splitting are shown in the figure below.

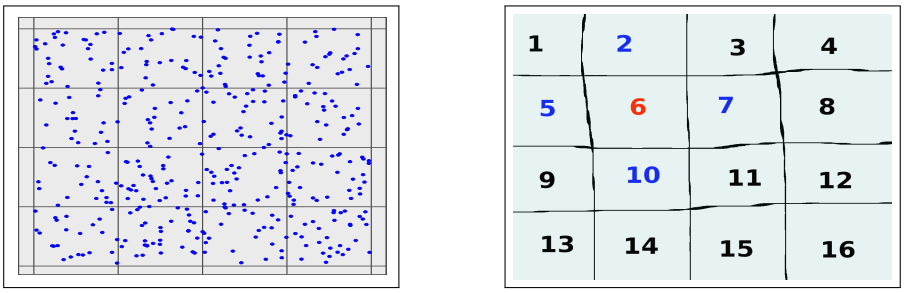

FIGURE 1.4: (left) A partitioned field, (right) generated lattice after spitting.

As shown in the first image of figure (1.4) the large field is divided into 16 small blocks. The picture in right is the lattice induced by splitting where the red colored block is connected with blue colored blocks. This blocking is done artificially and the respective lattice is very regular. However in the real field the lattice is not regular as this one rather they are similar as the right panel of figure (1.3).

Let us denote the i^{th} block estimate by $\hat{\theta}_i,\ i = 1, \ldots, K$, where K is the number of blocks. Then according to divide and conquer paradigm the global estimator is obtained as,

$$\hat{\theta}_{global} = \bigcup_{i=1}^{K} \omega_i \hat{\theta}_i \tag{1.3}$$

where ω_i is the block specific weight and $\bigcup$ meaning combination. The form of ω_i and $\bigcup$ varies depends on the methods. For example, ω_i is the inverse of prediction error and $\bigcup$ is the summation for spatial subsemble estimator.

This is the general form of global estimator under divide and conquer framework in both spatial and not spatial settings. Many proposals came out during last couple of decades for dividing and recombine the local results. Most but not all the methods aim to handle the big data problem. Some of the approaches apply combining technique in the meta-analysis setting. Most of the proposal aimed to achieve computation efficiency, scalability and drawing statistical inference(see for example, Liang *et al.* (2013), Yang *et al.* (2007), Dean and Ghemawat (2008), Liu *et al.* (2018), Barbian and Assunção (2017), Bickel *et al.* (2012), Zhou and Song (2017), Jordan *et al.* (2013), Guhaniyogi and Banerjee (2018)).

The significant proposal includes but not least simple averaging, weighted averaging of local estimates, combining functions, iterative updating. In the `farmer` approach model-based combining is proposed.

In the next chapter we describe the mathematical details, implementation and other necessary adjustments of the `farmer` approach.

Chapter 2

The farmer approach: theory

2.1 Introduction

In this chapter we discuss the mathematics behind `farmer` approach, derivation of the estimator, their standard error and necessary adjustments. We derive the results considering the geostatistical model as an example model however for any large data model, the farmer approach can be applied following the proper procedure. The following sections sequentially describe the entire approach in a sequential manner.

The first step of the approach is to split the data which is similar to the *MapReduce* scheme. The next section focus on the splitting procedure proposed in `farmer` approach.

2.2 Splitting the data and estimation in block

This is a general divide and conquer framework where we split the data in the first step. An example of splitting is shown in figure (2.1). The splitting mechanism of the entire field into sub-fields may vary based on the situation. When the locations are uniformly distributed over the whole region or if the study variable is some natural phenomena that do not depend on the region-specific social policies, that is they are free from the social effects, such as measurement of carbon monoxide or dioxide, the region is split into blocks in rectangular shape without considering the natural boundaries. However, if there are meaningful natural boundaries with a good number of locations then naturally generated blocks can be considered. For example, the United States has the state boundary which can be considered as splitting rule for any type of variable over the US. On the other hand, if the study variables have a strong dependence on social or country policies we propose to split the region keeping the natural boundaries unbroken. An example could be the epidemiological problem where the disease status

outbreaks strongly depend on the country's health system. In these types of cases, random rectangular splitting could mislead the inference. If some parts of a region enter into the adjacent region with completely different scenarios, this can affect the estimates and their error in blocks. Also, as mentioned earlier that the shape and size of the blocks influence the inference therefore care should be taken in splitting the data. As an example, consider the African river-blindness data where we split the data keeping the national boundaries unbroken. Rather we split the big country into parts and smaller countries we merged together. Depending on the splitting mechanism the induced lattice could be regular or irregular.

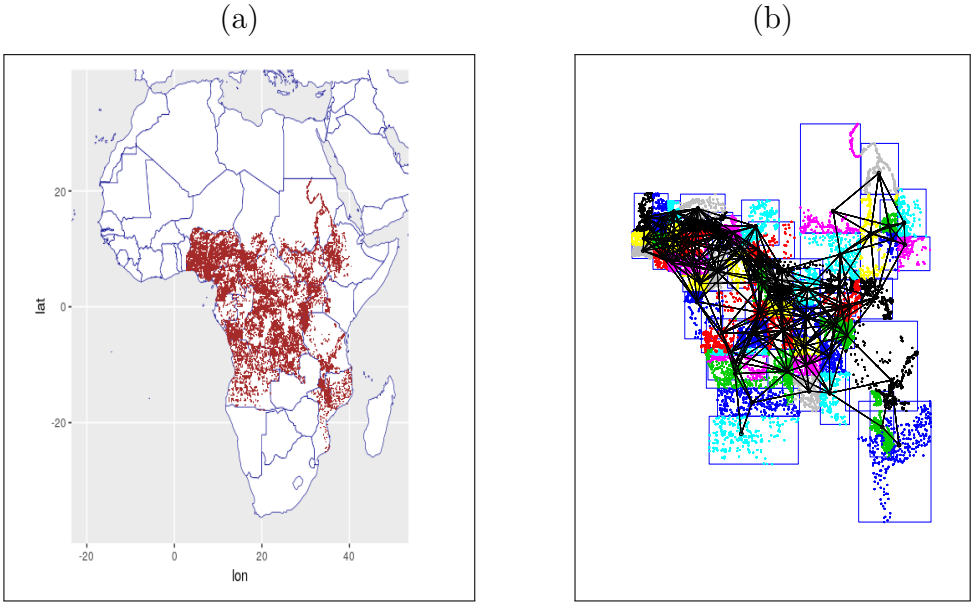

FIGURE 2.1: The locations of river-blindness data in Africa before (a) and after (b) splitting.

Once the data is divided we denote the data at block i as $z_i = \{y_i, X_i, s_i\}$ where y_i is the vector if response, X_i is the design matrix and s_i is the set of locations at that block. The subscript i takes the values $1, 2, \ldots, K$ for K mutually exclusive and exhaustive blocks. As we have defined the notations and parameters of the spatial process in the previous chapter, the process is defined by a set of parameters $\theta = (\mu, \sigma^2, \phi, \tau^2)$. We then estimate the model parameters and their variances at each block based on data z_i using some method, such as maximum likelihood. We denote the block summaries by the pair,

$$\{\hat{\theta}_i, V(\hat{\theta}_i)\}; \quad i = 1, 2, \ldots, K$$

The idea is to combine the local estimators even they are weaker. We propose to combine the block summaries using fixed and random effect meta analysis models. To do that we consider i^{th} block summary is the outcomes of a single study in meta analysis setting.

In our case we have multiple parameters in each block therefore we have to consider the multivariate meta analysis model.

In the next section, we describe how meta analysis model is used to obtain the `farmer` estimators and the related assumptions.

2.3 The `farmer` estimators

Meta analysis measures the common effect of some intervention programs based on multiple studies. This statistical technique generates very strong evidence in favor or disfavor of a treatment. This is now being widely used in different fields such as public health, psychology, medicine, and the social sciences. Long since there are two families of statistical models for performing meta analysis, the fixed and random effect models. The fixed effect model assumes the homogeneity of unknown effect across studies while the other assumes that the effect parameter is a random sample from a population (for details see Hedges and Vevea (1998)). However, there is some advance method for allowing heterogeneity in the meta analysis for multivariate case (see Liu *et al.* (2015)). Based on these two families of models we propose to pool the block estimates imagining each block estimate is the effect measure of a single study.

2.3.1 The fixed effect meta (`fem`) estimator

The fixed effect pooling assumes the following working model

$$\hat{\theta}_i = \theta + \epsilon_i; \quad i = 1, 2, \ldots, K; \quad \hat{\theta}_i \in \mathbb{R}^p \tag{2.1}$$

where $\hat{\theta}_i$ is the block estimates, θ is the parameter vector that defines the process, the vector of random errors ϵ_i's are assumed independent.

We will make the following further assumption,

- **A1.0** *the block errors,* $\epsilon_i \sim N_p(0, \Gamma_i) \Longrightarrow \hat{\theta}_i \sim N_p(\theta, \Gamma_i)$ and $\Gamma_i = V(\hat{\theta}_i)$ *where, p is the number of parameters to be estimated in the original model.*

However this is not always the case that the estimates are normally distributed, specially the estimate of the parameters (σ^2, ϕ, τ^2). This is a challenge. To overcome this challenge log transformation of parameters is a possible solution. The another useful solution is to adopt the Box-Cox transformation of the local estimates.

Then the fixed effect estimation procedure assumes the true model for $\hat{\theta} = (\hat{\theta}_1^\top, \hat{\theta}_2^\top, \ldots, \hat{\theta}_K^\top)$ is,

$$\hat{\theta} \sim N_{Kp}(1_K \otimes \theta, \Gamma), \tag{2.2}$$

where 1_K is a K-dimensional vector of ones, Γ is the block diagonal $Kp \times Kp$ covariance matrix with each block Γ_i is the $p \times p$ non spatial covariance matrix of the block estimates.

The fixed effect meta estimator of θ is the generalized least square solution of the model (2.1). The estimator is,

$$\hat{\theta}_{\texttt{fem}} = \{D^\top \Gamma^{-1} D\}^{-1} D^\top \Gamma^{-1} \hat{\theta}, \tag{2.3}$$

where $D = 1_k \otimes I_p$, with I_p the identity matrix of dimension p, Γ^{-1} is the inverse of Γ. To obtain the estimate (2.3) we need to estimate Γ which is a block diagonal matrix. Under the assumption **A1.0** the fixed effect meta estimator estimator is the maximum likelihood estimator for the model (2.1).

A simpler version of the estimator is obtained with equal weight for each block, that is if we replace $\Gamma_i = \gamma$, where, γ is a $p \times p$ common covariance matrix of the block estimates at every block, that is considering every block contribute equally which is statistically very naive. The simpler version is the block average which is obtained as,

$$\hat{\theta}_{\texttt{bar}} = \{D^\top \tilde{\Gamma}^{-1} D\}^{-1} D^\top \tilde{\Gamma}^{-1} \hat{\theta}, \tag{2.4}$$

where $\tilde{\Gamma}$ is a block diagonal matrix with common i^{th} diagonal element γ. The variance of $\hat{\theta}_{\texttt{fem}}$ is obtained as below,

$$\text{var}(\hat{\theta}_{\texttt{fem}}) = \{D^\top \Gamma^{-1} D\}^{-1} D^\top \Gamma^{-1} \text{var}(\hat{\theta}) \Gamma^{-1} D \{D^\top \Gamma^{-1} D\}^{-1} \tag{2.5}$$

If the model (2.1) is true, that is the working model coincides the true model then the variances reduced to,

$$\text{var}(\hat{\theta}_{\texttt{fem}}) = \{D^\top \Gamma^{-1} D\}^{-1}. \tag{2.6}$$

The two versions of variance for the estimator $\hat{\theta}_{\texttt{bar}}$ can be obtained as well by replacing $\Gamma = \tilde{\Gamma}$ in (2.5) and (2.6).

Now, if the assumption **A1.0** is correct then (2.5) and (2.6) should be same. However, this is better to estimate the variance (2.5) using some method. One way of doing this is to employ spatial heteroscedastic and auto-correlation consistent (SHAC) estimation procedure. We will describe the SHAC estimation procedure later section.

The next subsection describes the procedure of obtaining random effect meta estimator and its variances. We have employed the random effect meta analysis model which allows us to take into account the between block dependence.

2.3.2 The random effect meta (rem) estimator

To account for the between block dependence we propose to combine the local estimators through a random effect meta analysis model which assumes that the local estimators differ from the true parameter value by a random quantity that smoothly varies across the blocks. The assumed working model is,

$$\hat{\theta}_i = \theta + \eta_i + \epsilon_i, \tag{2.7}$$

where η_i are zero-mean random effects designed in a way to describe the spatial variation of the local estimators ($i = 1, \ldots, K$) and ϵ_i accounts for within block variation, also can be termed as estimation error.

In addition to the assumptions made for fixed effect meta estimation we make the following further assumptions,

- **A2.0** *the random effects η_i are independent from the estimation errors ϵ_i.*
- **A2.1** *the random effects $\eta \sim N_{Kp}(0, \Omega)$, where $\eta = \{\eta_i; i = 1, \ldots, K\}$.*

Hence, the random effect meta-analysis model assumes that,

$$\hat{\theta} \sim N_{Kp}(1_K \otimes \theta, \Gamma + \Omega), \tag{2.8}$$

where Ω is the covariance matrix of the random effects which represent the between block dependence and Γ is the matrix of withing block dependence of the block estimates. In the case of random effect meta estimation the covariance matrix $(\Gamma + \Omega)$ is not block diagonal anymore. Generalized least square is the classical approach usually adopted for obtaining the random effect meta estimator in meta analysis settings. Therefore, in the `farmer` approach the random effect meta estimator, $\hat{\theta}_{\texttt{rem}}$ is obtained following the generalized least square approach. The estimator is obtained as,

$$\hat{\theta}_{\texttt{rem}} = \{D^\top(\Gamma+\Omega)^{-1}D\}^{-1}D^\top(\Gamma+\Omega)^{-1}\hat{\theta}, \tag{2.9}$$

where D is defined as in previous section and the value of Γ is obtained as described in previous subsection. The Ω needs to be estimated. The estimation procedure will be discussed in the next subsection. This is easy to notice that if the between block covariance Ω become null then the $\hat{\theta}_{\texttt{rem}}$ reduces to $\hat{\theta}_{\texttt{fem}}$. Also, the $\hat{\theta}_{\texttt{rem}}$ reduces to $\hat{\theta}_{\texttt{bar}}$ if we replace $(\Gamma+\Omega) = \tilde{\Gamma}$.

In the similar way as described for fixed effect meta estimation the variance of the random-effects meta estimator is

$$\text{var}(\hat{\theta}_{\texttt{rem}}) = \{D^\top(\Gamma+\Omega)^{-1}D\}^{-1}D^\top(\Gamma+\Omega)^{-1}\text{var}(\hat{\theta})(\Gamma+\Omega)^{-1}D\{D^\top(\Gamma+\Omega)^{-1}D\}^{-1}. \tag{2.10}$$

If the working model (2.7) coincides the true model in (2.8) then the variance of the `farmer rem` estimator reduces to,

$$\text{var}(\hat{\theta}_{\texttt{rem}}) = \{D^\top(\Gamma+\Omega)^{-1}D\}^{-1} \tag{2.11}$$

Again, replacing the estimate for Ω we will obtain the `farmer rem` estimator and related variances.

Both the fixed and random effect meta analysis models provide us the analytical solution for `farmer` estimators and their variances. Specially, for `farmer` $\hat{\theta}_{\texttt{fem}}$ estimator all the required quantities are readily available from the block estimates except the term $\text{var}(\hat{\theta})$ in the variance formula. The estimation procedure of this quantity will be discussed in the later section. However, for the `farmer rem` estimator we have to specify the covariance for the random effect which is Ω.

Splitting of large data domain into mutually exclusive and exhaustive blocks leave us the lattice. The generated lattice can be regular or irregular based on the splitting mechanism. There is huge literature for modeling the lattice data. The seminal article of Besag (1974) has opened the windows for dealing with Gaussian and non-Gaussian lattice data. Later Gelfand and Vounatsou (2003), Jin *et al.* (2005) have done further developments in the multivariate case of lattice data and suggested of multivariate conditional autoregressive (MCAR) models and the generalized version of MCAR. In our problem we have multiple parameters in each block therefore we fit the MCAR model for estimating the random effect covariance parameter Ω. In the following subsection

estimation procedure for Ω is detailed out.

2.3.3 The specification of Ω

As we mentioned earlier partitioning the domain into mutually exclusive and exhaustive blocks produces either regular or irregular lattice based on the splitting rule. In a lattice, the blocks share their borders with their neighbors' blocks. The blocks that share a common border are correlated and if there is no common border between two blocks they are uncorrelated. This is the simple intuitive idea behind the conditional autoregressive (CAR) model. The model is formulated using the hierarchical approach. Let us consider the figure (2.2) which is an example of regular lattice consists of 16 areal units induced by splitting. The blocks are usually called an areal unit in spatial literature. Here, the block 6 (in red) shares it's border with blocks $(2, 5, 7, 10)$ therefore the block 6 is associated with blocks $(2, 5, 7, 10)$, also block 4 shares border with blocks $(3, 8)$ and similarly they are associated. This is called the first level of correlation. For simplicity, we will consider the first level of correlation only in this thesis.

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

FIGURE 2.2: Blocking generated lattice

For the convenience of readers we first discuss briefly the formulation of $MCAR$ model starting from univariate case that is the formulation of CAR model. Let us consider η_i is a random variable observed at K areal units, that is in K blocks in our setting. Then under MRF assumptions and following the Besag (1974)'s formulation, the K full conditionals are defined as,

$$p(\eta_i|\eta_j, i \neq j, \lambda_i) = N\left(\nu \sum_{i\sim j} b_{ij}\eta_j, \lambda_i\right), \quad i, j = 1, 2, \ldots, K, \tag{2.12}$$

where $i \sim j$ means that the unit i is a neighbor of unit j, λ_i is the conditional variance of η_i, b_{ij} conveys the spatial dependence between the i^{th} and j^{th} location. From Hammersley-Clifford Theorem and Brook's Lemma (see Banerjee *et al.* (2014), section 4.2) the joint distribution of $\eta = (\eta_1, \eta_2, \ldots, \eta_K)$ can be obtained as,

$$\eta \sim N_K\left(\mathbf{0}, [D_\lambda(I - \nu B)^{-1}]\right), \tag{2.13}$$

which is a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathrm{D}_\lambda(I - \nu B)^{-1}$. Where $B = \{b_{ij}\}$ is a $K \times K$ matrix with element $b_{ii} = 0$, D_λ is $K \times K$ diagonal matrix with non-zero entries λ_i. The parameter ν is a smoothness parameter which controls the spatial dependence among the blocks, $\nu = 0$ implies an independent model however $\nu = 1$ does not imply a completely dependent model rather the distribution becomes improper. This parameter lies inside 0 and 1. In the analysis of lattice data the adjacency matrix W is an important matrix which represent the neighborhood structure among the blocks. There are many ways to define the matrix W however the simple definition is,

$$w_{ij} = \begin{cases} 1 & \text{if i and j share a common border} \\ 0 & \text{otherwise .} \end{cases} \tag{2.14}$$

This definition is not always recommended for irregular lattice. In that case, some weighting scheme based on distance is recommended (see for example Banerjee *et al.* (2014)). Once the W is defined the we set $b_{ij} = w_{ij}/w_{i+}$, where, $w_{i+} = \sum_j w_{ij}$. If $\lambda_i = \lambda/w_{i+}$ holds, then we have,

$$\mathrm{D}_\lambda^{-1}(I - \nu B) = \frac{1}{\lambda}(D_W - \nu W),$$

where λ is the common variance of the random variable η_i across the blocks and $D_W = \mathrm{diag}(w_{1+}, w_{2+}, \ldots, w_{K+})$. This model is denoted by $CAR(\nu, \lambda)$. Based on the values of ν there could be variants of the CAR model. Then the variance in the model (2.13) becomes $\lambda(D_W - \nu W)^{-1}$.

Now, let us consider that $\eta^\top = (\eta_1^\top, \eta_2^\top, \ldots, \eta_K^\top)$ where each η_i is a $p \times 1$ vector. The the full conditional is defined as,

$$p(\eta_i|\eta_j, i \neq j, \Lambda_i) = N_p\left(\nu \sum_{i\sim j} b_{ij}\eta_j, \Lambda_i\right), \quad i, j = 1, 2, \ldots, K, \tag{2.15}$$

where the smoothness parameter ν is common for all p elements in η_i, also, the b_{ij}'s are same for all the elements in η_i and Λ_i is the $p \times p$ covariance matrix of vector η_i. Then the univariate CAR model is not suitable rather we have multivariate CAR ($MCAR$) models. Following the similar formulation, we have the $MCAR(\nu, \Lambda)$ model as,

$$\eta \sim N_{Kp}\left(\mathbf{0}, [(D_W - \nu W)^{-1} \otimes \Lambda]\right), \tag{2.16}$$

where $\Lambda_i = \Lambda/w_{ij}$ and Λ is the $p \times p$ common covariance matrix of vector η_i, W and D_W are now $Kp \times Kp$ matrices. Under the assumption in (2.8) and formula (2.16) the variance of vector η is given by,

$$\Omega = (D_W - \nu W)^{-1} \otimes \Lambda, \tag{2.17}$$

where every components has same interpretation as mentioned before.

According to the model (2.8) the $\text{var}(\hat{\theta}) = \Gamma + \Omega$, therefore the log-likelihood function ignoring the constant terms is,

$$\begin{aligned} l(\theta, \nu, \Lambda) = &-\frac{1}{2}\log|\Gamma + (D_W - \nu W)^{-1} \otimes \Lambda| \\ &-\frac{1}{2}(\hat{\theta} - 1_K \otimes \theta)^\top(\Gamma + (D_W - \nu W)^{-1} \otimes \Lambda)^{-1}(\hat{\theta} - 1_K \otimes \theta). \end{aligned} \tag{2.18}$$

This is the likelihood function for all parameters. However, we have the analytical estimator for θ which is a function of ν, Λ through Ω as,

$$\hat{\theta}_{\texttt{rem}} = \{D^\top(\Gamma + \Omega)^{-1}D\}^{-1}D^\top(\Gamma + \Omega)^{-1}\hat{\theta}.$$

Therefore we need to estimate the last two parameters ν, Λ involved in Ω only. We can express estimator of θ as $\hat{\theta}_{\texttt{rem}}(\nu, \Lambda)$ and replace this into (2.18) to get the profile likelihood for (ν, Λ) as,

$$l^{\star}(\nu,\Lambda) = -\frac{1}{2}\log|\Gamma + (D_W - \nu W)^{-1} \otimes \Lambda| \\ -\frac{1}{2}(\hat{\theta} - 1_K \otimes \hat{\theta}_{\texttt{rem}}(\nu,\Lambda))^T(\Gamma + (D_W - \nu W)^{-1} \otimes \Lambda)^{-1}(\hat{\theta} - 1_K \otimes \hat{\theta}_{\texttt{rem}}(\nu,\Lambda)). \tag{2.19}$$

We have to maximize the (2.19) for estimating (ν, Λ), where Γ is known from block estimates as described before. The `farmer` $\hat{\theta}_{\texttt{rem}}$ estimator for θ is then be defined by plugging $(\hat{\nu}, \hat{\Lambda})$ in (2.9).

At this point, there are several possibilities for specifying the matrix Λ. Firstly, we can set all the $p + p(p-1)/2$ parameters included in the matrix as unknown; second, the off-diagonal block between regression parameters and spatial parameters can be set equal to zero; thirdly, fixing all the elements of Λ by either empirical covariance matrix from the block estimates or average of block covariance matrix $V(\hat{\theta}_i)$. The dimension of Λ depends on the number of parameters to be estimated in the model. If there are many covariates included in the model then the dimension of Λ explodes. Considering this issue we have adopted the last option. This makes the estimation of the $MCAR(\nu, \Lambda)$ model computationally efficient and we need to estimate a single parameter only. At this stage, we have a further modification of the profile likelihood which is,

$$l^{\star\star}(\nu) = -\frac{1}{2}\log|\Gamma + (D_W - \nu W)^{-1} \otimes \tilde{\Lambda}| \\ -\frac{1}{2}(\hat{\theta} - 1_K \otimes \hat{\theta}_{\texttt{rem}}(\nu,\tilde{\Lambda}))^T(\Gamma + (D_W - \nu W)^{-1} \otimes \tilde{\Lambda})^{-1}(\hat{\theta} - 1_K \otimes \hat{\theta}_{\texttt{rem}}(\nu,\tilde{\Lambda})). \tag{2.20}$$

where

$$\tilde{\Lambda} = \frac{1}{K}\sum_{i=1}^{K} V(\hat{\theta}_i).$$

Now, the problem is simplified, computation is feasible and the number of parameters is independent of the number of blocks and p.

Since we are not sure that the working model coincides with the true model we think this is not wise to use the simplified formula for variance estimation. In this regards it is necessary to consider the term $\text{var}(\hat{\theta})$ in the variance formulas. As said earlier, we propose to adopt the non-parametric approach *Spatial Heteroscadastic and Autocorrelation Consistent (SHAC)* estimation procedure for the required estimation. Through SHAC we estimate the variances for `farmer fem` and `rem` estimators. The next

section is focused on describing the SHAC estimation procedure for variance estimation.

2.3.4 Estimation of variances for $\hat{\theta}_{\texttt{bar}}$, $\hat{\theta}_{\texttt{fem}}$, and $\hat{\theta}_{\texttt{rem}}$

Heteroscedasticity and autocorrelation consistent (HAC) estimator is a non-parametric estimator that can be used for many sample statistics. This method is more frequent in the time series and econometric literature. Grenander and Rosenblatt (1957) in their book described the nut and bolts of the estimator which is a classical reference to go through. For recent updates please see the latest version of the book (Grenander and Rosenblatt (2008)). Priestley (1964) for the first time introduced the HAC estimator in the spatial context for estimating the spectral densities of stationary random fields. Further development on this estimator is done by Kelejian and Prucha (2007) in the spatial setting and the name Spatial HAC (SHAC) is introduced by them. They establish the non-parametric SHAC based on estimated disturbances for estimating the variance-covariance matrix for the sample moments and demonstrated the consistency of the estimator under mild conditions. We have adopted the SHAC estimator defined by Kelejian and Prucha (2007) which is suitably fit to our problem.

To explain the methodology we consider the following example of specific regression model:

$$y = X\beta + e \tag{2.21}$$

where y is the vector of n data points associated with n locations, X is the $n \times p$ design matrix and β is the vector of regression parameters associated with the p covariates, n is the number of spatial units in the sample. The OLS estimator of regression coefficient, $\hat{\beta}$ is defined as,

$$\hat{\beta} = (X^\top X)^{-1} X^\top y \tag{2.22}$$

Now, the question is to estimate the covariance matrix of $\hat{\beta}$ considering the spatial association between the points and correlation structure is unknown. In this situation the variance covariance of $\hat{\beta}$ is defined based on SHAC procedure as,

$$\begin{aligned} \Psi &= \text{var}\left(\sqrt{n}(\hat{\beta} - \beta)\right) \\ &= n^{-1} X^\top \Sigma_e X \end{aligned} \tag{2.23}$$

where Σ_e is the covariance of error. Then the SHAC estimator of Ψ is given by,

$$\hat{\Psi} = n^{-1} X^\top \hat{e}^\top \hat{e} \circ K(d/d_n) X, \tag{2.24}$$

where $\circ$ means element by element multiplication, $K(\cdot)$ is the kernel function, $\hat{e}$ is the estimtaed error vector, d is the distance matrix, d_n is the scaling factor of distance. The scaling factor d_n has similar type of interpretation of the spatial scale parameter ϕ however depends on the kernel function used. The large value of d_n means that the distant blocks are also correlated on the other hand the smaller value indicate that the closer blocks are only correlated. When d_n takes the value equal maximum distance then all the blocks are correlated and when it takes the value equal 1 then the blocks with distance less than or equal 1 unit are correlated others not. The dimension of $\hat{\Psi}$ is $p \times p$, p is the dimension of β. This is the formulation in Kelejian and Prucha (2007).

The authors have suggested several different kernel functions. The example of some kernels is mentioned below. The ij^{th} element of which is defined as,

Truncated:

$$K_{TR}(d_{ij}) = \begin{cases} 1, & \text{if } |d_{ij}| \leqslant 1 \\ 0, & \text{otherwise} , \end{cases}$$

Bartlett:

$$K_{BT}(d_{ij}) = \begin{cases} 1 - |d_{ij}|, & \text{if } |d_{ij}| \leqslant 1 \\ 0, & \text{otherwise} , \end{cases}$$

Parzen:

$$K_{PR}(d_{ij}) = \begin{cases} 1 - 6d_{ij}^2 + 6|d_{ij}|^3, & \text{if } 0 \leqslant |d_{ij}| \leqslant 1/2 \\ 2(1 - |d_{ij}|)^3, & \text{if } 1/2 \leqslant |d_{ij}| \leqslant 1 \\ 0, & \text{otherwise} , \end{cases}$$

Tukey-Hanning:

$$K_{TH}(d_{ij}) = \begin{cases} (1 + \cos(\pi d_{ij}))/2, & \text{if } |d_{ij}| \leqslant 1 \\ 0, & \text{otherwise} , \end{cases}$$

Quadratic Spectral:

$$K_{QS}(d_{ij}) = \frac{25}{12\pi^2 d_{ij}^2} \Big(\frac{\sin(6\pi d_{ij}/5)}{6\pi d_{ij}/5} - \cos(6\pi d_{ij}/5) \Big).$$

Now, this is clear that the $\hat{\theta}_{\texttt{bar}}$, $\hat{\theta}_{\texttt{fem}}$ and $\hat{\theta}_{\texttt{rem}}$ have the similar structure with (2.22). We can easily replace the X by $(D^\top\Gamma^{-1}D)^{-1}D^\top\Gamma^{-1}$ for `fem` and $(D^\top(\Gamma+\hat{\Omega})^{-1}D)^{-1}D^\top(\Gamma+\hat{\Omega})^{-1}$ for `rem`. Also, the $\hat{e}$ is replaced by $(\hat{\theta}-\hat{\theta}_{\texttt{farmer}})$. Replacing this we obtain the SHAC estimators for the variances as below:

$$\hat{\text{var}}(\hat{\theta}_{\texttt{farmer}}) = A\{(\hat{\theta}-\hat{\theta}_{\texttt{farmer}})^\top(\hat{\theta}-\hat{\theta}_{\texttt{farmer}}) \circ (K(d/d_n)\otimes \mathbb{1})\}A^\top \tag{2.25}$$

where $A = (D^\top(\Gamma+\hat{\Omega})^{-1}D)^{-1}D^\top(\Gamma+\hat{\Omega})^{-1}$ and $A = (D^\top\Gamma^{-1}D)^{-1}D^\top\Gamma^{-1}$ for `farmer` `rem` and `fem` respectively and $\mathbb{1}$ is $p \times p$ matrix of 1.

To get the three versions of farmer estimate `bar`, `fem` and `rem` and their variances we now have all the requirements. However, to follow the process properly, we have organized the entire process in the form of an algorithm. The farmer algorithm is presented in the following section.

2.4 The `farmer` algorithm

To implement the `farmer` approach the following algorithm need to be followed up. The simple and intuitive farmer algorithm are elaborated in a step-wise manner to follow through easily.

Algorithm 1 `farmer algorithm`

- Partition the data set, $z = (y, X, s)$ into K non-overlapping subsets and label the subsets from $1, 2, \ldots, K$ each of which are called as *"block"*. The set (y, X, s) are respectively the observed realization of study variable, a possible vector of covariates, and set of locations. Denote the data in block i as z_i with $z_i = (y_i, X_i, s_i)$.
- Obtain the block estimates $\hat{\theta}_1, \ldots \hat{\theta}_K$ using some method (for example maximum likelihood) from the data $z_i; i = 1, 2, \ldots, K$ and calculate $\theta^{(0)} = \sum_1^K \hat{\theta}_i/K$.
- Calculate $V(\hat{\theta}_i);\ i = 1, \ldots, K$. Also, set $\tilde{\Lambda} = \sum_1^K V(\hat{\theta}_i)/K$.
- Minimize the negative log likelihood (2.20) to obtain $\hat{\nu}$ and then calculate $\hat{\Omega}$.
- Calculate three versions of farmer estimators $\hat{\theta}_{\texttt{bar}}$, $\hat{\theta}_{\texttt{fem}}$, and $\hat{\theta}_{\texttt{rem}}$ using the formulas (2.4), (2.3) and (2.9) respectively.
- Obtain variances for different farmer estimators using equation (2.25). The simplified versions of the variances are obtained by the formula defined in (2.6) and (2.11).

2.5 Likelihood and Fisher information matrix for Gaussian random fields

We explain the parameter estimation and finding the respective variance in block using maximum likelihood approach as an example. However, the `farmer` approach is a general platform where other method of estimation can be employed too. Let us consider $Y(s)$ is a Gaussian random field observed at location $s \in \mathbb{R}^2$. The simple geostatistical model for $Y(s)$ is,

$$Y(s) = X(s)^\top \beta + S(s) + e(s) \tag{2.26}$$

where $X(s)$ are the possible vector of known covariates associated with locations s, and β is a vector of parameters associated with the covariates. $S(s)$ has a zero mean Gaussian process which represent the spatial variation in the process also known as latent process. If there is n locations then the latent vector $S = (S(s_1), \ldots, S(s_n))$ has an n dimensional multivariate normal distribution which is $N(0, \sigma^2 R(\phi))$, where σ^2 is

the variance parameter, $R(\phi)$ is the $n \times n$ correlation matrix with scale parameter ϕ. The nugget vector $e = (e(s_1), \ldots, e(s_n))$ assumed to have $N(0, \tau^2 I)$, where τ^2 is called the nugget variance and I is the identity matrix of order n. Then the variance covariance matrix of $Y = (Y(s_1), \ldots, Y(s_n))$ is $\Sigma = \sigma^2 R(\phi) + \tau^2 I$. The log likelihood function for the model under consideration is then be expressed as,

$$l(\theta, y) = -\frac{1}{2} n \log(2\pi) - \frac{1}{2} \log|\Sigma| - \frac{1}{2}(y - X\beta)^\top \Sigma^{-1}(y - X\beta) \tag{2.27}$$

where y is the vector of realizations of Y, $\theta = (\beta, \sigma^2, \phi, \tau^2)$ is the vector of parameters to be estimated, the last three parameters are termed commonly as spatial dependence parameters. Let us divide the parameter set into two subsets $\theta = (\beta, \varrho)$, where $\varrho = (\sigma^2, \phi, \tau^2)$. For any fixed value ϱ, say, ϱ_0, the MLE for β can be obtained maximizing (2.27). The MLE for β is obtained as,

$$\hat{\beta} = \left(X^\top \Sigma^{-1}(\varrho_0) X\right)^{-1} X^\top \Sigma^{-1}(\varrho_0) y.$$

Unfortunately there is no closed form solution for the parameters ϱ. Therefore, we have to estimate ϱ by numerically maximizing the profile likelihood. Replacing β by $\hat{\beta}$ we obtain the profile likelihood function for ϱ as,

$$l_p(\varrho, y) = -\frac{1}{2} n \log(2\pi) - \frac{1}{2} \log|\Sigma(\varrho)| - \frac{1}{2} y^\top H^{-1}(\varrho) y, \tag{2.28}$$

where $H(\varrho) = \Sigma^{-1}(\varrho) - \Sigma^{-1}(\varrho) X (X^\top \Sigma^{-1}(\varrho) X)^{-1} X^\top \Sigma^{-1}(\varrho)$, and the estimate of β is updated $\hat{\beta} = \left(X^\top \Sigma^{-1}(\hat{\varrho}) X\right)^{-1} X^\top \Sigma^{-1}(\hat{\varrho}) y$.

The covariance matrix of the estimate $\hat{\theta}$ can be obtained inverting the Fisher information matrix. The Fisher information matrix based on (2.27) is obtained as,

$$I(\theta) = \begin{pmatrix} I_{\beta\beta} & I_{\beta\sigma^2} & I_{\beta\phi} & I_{\beta\tau^2} \\ I_{\sigma^2\beta} & I_{\sigma^2\sigma^2} & I_{\sigma^2\phi} & I_{\sigma^2\tau^2} \\ I_{\phi\beta} & I_{\phi\sigma^2} & I_{\phi\phi} & I_{\phi\tau^2} \\ I_{\tau^2\beta} & I_{\tau^2\sigma^2} & I_{\tau^2\phi} & I_{\tau^2\tau^2} \end{pmatrix} \tag{2.29}$$

where the elements are defined as,

$$
\begin{aligned}
I_{\beta\beta} &= X^{\top}\Sigma^{-1}X \\
I_{\sigma^2\sigma^2} &= \frac{1}{2}\text{tr}\left[\Sigma^{-1}R(\phi)\Sigma^{-1}R(\phi)\right] \\
I_{\sigma^2\phi} &= \frac{1}{2}\text{tr}\left[\Sigma^{-1}R(\phi)\Sigma^{-1}\sigma^2 R^{'}(\phi)\right] \\
I_{\sigma^2\tau^2} &= \frac{1}{2}\text{tr}\left[\Sigma^{-1}R(\phi)\Sigma^{-1}I\right] \\
I_{\phi\phi} &= \frac{1}{2}\text{tr}\left[\Sigma^{-1}\sigma^2 R(\phi)^{'}\Sigma^{-1}\sigma^2 R^{'}(\phi)\right] \\
I_{\phi\tau^2} &= \frac{1}{2}\text{tr}\left[\Sigma^{-1}\sigma^2 R^{'}(\phi)\Sigma^{-1}I\right] \\
I_{\tau^2\tau^2} &= \frac{1}{2}\text{tr}\left[\Sigma^{-1}I\Sigma^{-1}I\right] \\
I_{\beta\sigma^2} &= I_{\beta\phi} = I_{\beta\tau^2} = 0,
\end{aligned}
$$

where $R^{'}(\phi)$ the first derivatives of correlation matrix $R(\cdot)$ with respect to ϕ. We can obtain the $I(\hat{\theta})$ by replacing the θ with respective estimated value. We apply the ML estimation procedure based on above formulation to obtain the $\hat{\theta}_i$ and $I(\hat{\theta}_i)$ at every block. Then we obtain the $V(\hat{\theta}_i) = I^{-1}(\hat{\theta}_i)$.

This is to be noted that we have assumed the asymptotic normality of block estimates $\hat{\theta}_i$ when fixed and random effect meta analysis applied. However, this is not always the case, especially for the estimate of the spatial dependence parameters ϑ. This is a challenge. To overcome this challenge log transformation of these parameters is a possible solution. Another useful solution is to adopt the Box-Cox transformation of the local estimates. For this example, we have re-parameterized the block likelihood to obtain the log-transformed spatial parameters and respective information matrix. Which provides the closer approximation to the normality. This is a simple and intuitive solution to the problem. This transformation makes us free from constrained optimization that is all the parameters can take values over $-\infty$ *to* ∞.

In the following section, we have checked how `farmer` approach performs with a simple simulation example. Details are described in the section below.

2.6 A first performance checking of `farmer` algorithm

For assessing farmer algorithm's performance we have experimented. We have considered the model (2.26) with a single covariate. We have generated $n = 20,000$ non-regular locations uniformly over a range of $(0, 30) \times (0, 30)$. The true values are considered as

$\beta_0 = 1, \beta_1 = 1$. We have considered the exponential covariance model with parameters values $\sigma^2 = 1, \phi = 0.15, \tau^2 = 0$. The covariate is considered as known which comes from $N(0, 0.5^2)$. We have split the region into 80 mutually exclusive blocks of size ≈ 250 each. The likelihood is defined in a way to accommodate the log transferred spatial parameters. The proposed farmer approach is applied to this generated data set for obtaining the estimates of the model. We have repeated the experiment 200 times. The outcomes are presented in table (2.1) and figure (2.3). In the table, we have presented 10% trimmed mean.

Estimators	β_0	β_1	$\log(\sigma^2)$	$\log(\phi)$
true	1.00	1.00	0.00	-1.8971
bar	0.9998	0.9999	-0.0233	-1.9259
fem	0.9998	0.9999	-0.0233	-1.9259
rem	0.9998	0.9999	-0.0233	-1.9259

TABLE 2.1: The bar, fem, rem estimates from 200 experiments

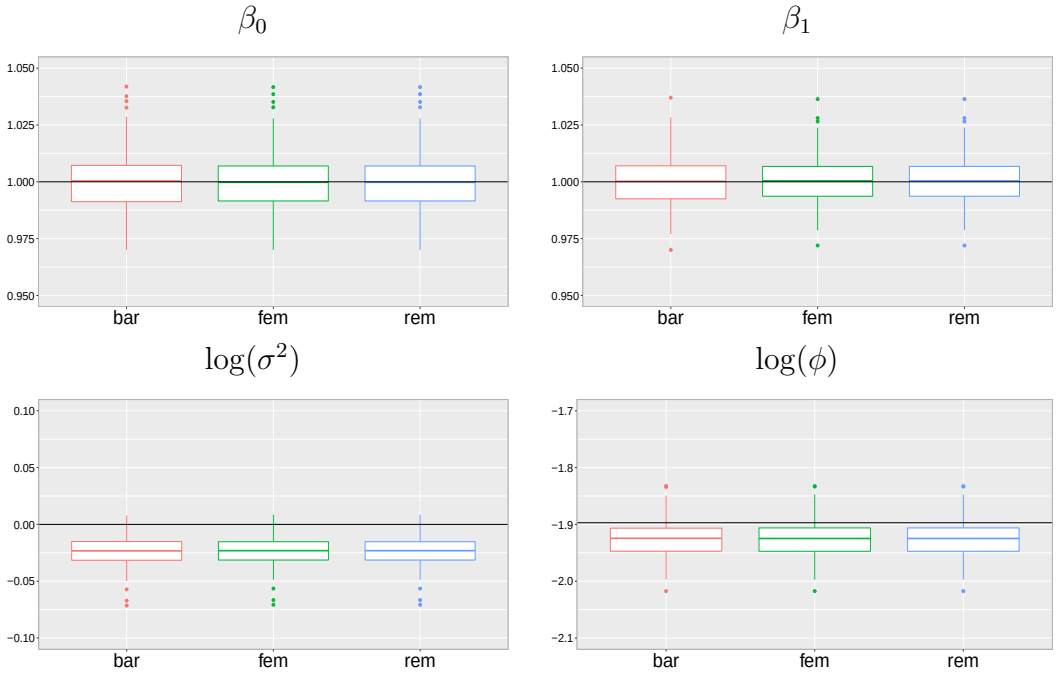

FIGURE 2.3: Boxplots of the bar, fem, rem, for betas, log sigma square and log phi.

Conducting the first set of simulation experiments we understood that the farmer approach is easy to apply and computationally efficient. The regression parameters are captured very well by the proposed approach which is the case for other approaches as well. The estimation time is also within the minutes' range, on average 15 minutes for each iteration using a personal computer of core i7, 2.7GHz processor, Ubuntu 18.04 OS. However, there is evidence of downward bias in spatial parameters (σ^2, ϕ). Therefore

we may say that using the farmer approach it is easy to catch the mean however for the variance part there is some disturbance. This downward bias in the maximum likelihood estimator is not new, some other authors also noted before (Viechtbauer (2005)). Also, in spatial setting Barbian and Assunção (2017) and Liang *et al.* (2013) have experienced underestimation of these parameters and instead they reported the ratio σ^2/ϕ of these two parameters and they left the problem unsolved.

Smith (2009) has detailed out the root causes of the downward bias of spatial dependence parameters. The *strong connectivity* is the main culprit for this type of bias. The authors mentioned several studies that experienced the same issue. That is the strongly connected or high-density weight matrics are the cause of happening this bias. The authors identified the cause however no remedial measure has been suggested. This is is still now an open question. Firth (1993) suggested a bias correction method in the likelihood. Further development has been done on this later by Kosmidis *et al.* (2017) who implemented the former idea for improving the accuracy of likelihood inference for random effect meta analysis and meta-regression.

In the next section, we describe the procedure of adjustment for bias correction for `farmer` estimators.

2.7 Bias reduction for `farmer` estimators

We propose further improvements of farmer estimators by employing this bias correction method at the block level.

In the `farmer` approach, the likelihood method is being applied in two stages, first, at the block level and second for estimating the MCAR model. Therefore the problem of bias can occur in both of the stages. Based on the source of bias the reduction strategy can be different. According to the first simulation results presented in figure (2.3) this is clear that every estimator such as `bar`, `fem` and `rem` have this problem of underestimation. It is clear that the second stage model is only applied for `rem` estimator and bias have occurred for every estimator. This is suggestive that the bias actually occurs when estimating in blocks. Therefore, we propose to apply the bias correction technique at the block levels.

Firth (1993) proposed a bias correction technique in likelihood estimation settings for independent data. In this method instead of reducing bias he has suggested to introduce some bias into the score function. This modification of score function is done based on simple triangle geometry (see figure 1 in Firth (1993)). Let us consider the log likelihood function $l(\theta)$ to be maximized for estimating the parameter θ. According to

the proposal for positive bias, the modified log likelihood function is,

$$l^{\star}(\theta) = l(\theta) + \frac{1}{2}\log|I(\theta)|, \tag{2.30}$$

where $|I(\theta)|^{\frac{1}{2}}$ is the Jeffreys invariant prior penalty function. Based on Firth (1993)'s works, Kosmidis *et al.* (2017) suggested more specific correction as mean bias correction and later Kyriakou *et al.* (2018) have updated it and proposed a median bias correction method for random effect meta analysis model. They proposed the modification of score function for downward bias of maximum likelihood estimators. Recently, Kosmidis *et al.* (2018) have proposed the mean and median bias reduction techniques for generalized linear models.

In our case, we have experienced downward bias too. Therefore we propose to apply the Kosmidis *et al.* (2017)'s technique to the log-likelihood function (2.27). and we obtain the mean bias-adjusted log-likelihood function,

$$l^{\star}(\theta, y) = -\frac{1}{2}n\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(y - X\beta)^{\top}\Sigma^{-1}(y - X\beta) - \frac{1}{2}\log|I_{\beta\beta}(\theta)|. \tag{2.31}$$

Maximizing the equation (2.31) for each block we obtain the mean bias corrected block estimates. The algorithm (1) will then be employed on mean bias corrected block estimates. The author showed that the inference based on penalized likelihood is same as the inference based on the usual likelihood. We will denote the bias reduced estimates by $\hat{\theta}^{\dagger}_{\texttt{bar}}, \hat{\theta}^{\dagger}_{\texttt{fem}},$ and $\hat{\theta}^{\dagger}_{\texttt{rem}}$.

To see the impact of the bias reduction method we have estimated bias reduced parameters on the same simulated data presented in the previous section. The comparative results are as presented below table and figure.

The output from 200 experiments are presented below in table (2.2) and in figure (2.4).

type	Estimators	β_0	β_1	$\log(\sigma^2)$	$\log(\phi)$
	true	1.00	1.00	0.00	-1.8971
No BC	bar	0.9998	0.9999	-0.0233	-1.9259
	fem	0.9998	0.9999	-0.0233	-1.9259
	rem	0.9998	0.9999	-0.0233	-1.9259
Mean BC	bar	0.9998	0.9999	-0.0066	-1.9057
	fem	0.9998	0.9999	-0.0066	-1.9057
	rem	0.9998	0.9999	-0.0066	-1.9057

TABLE 2.2: The bar, fem, rem estimates and their bias corrected forms from 200 experiments

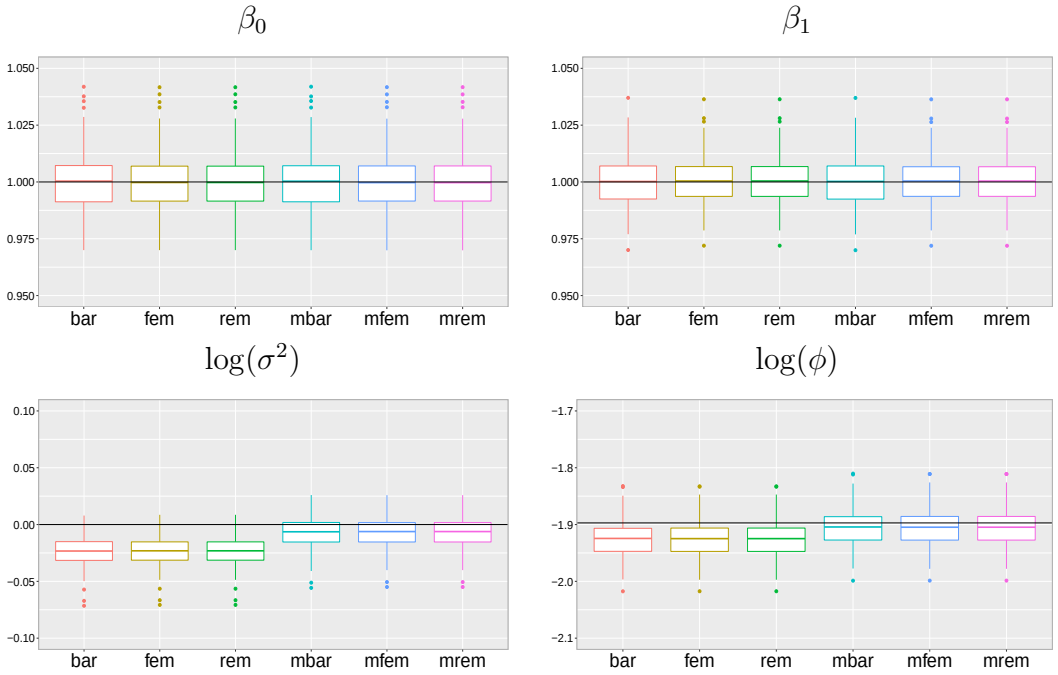

FIGURE 2.4: Boxplots of the bar, fem, rem, for betas, sigma square, and phi. In each panel, the first three boxes are bias-uncorrected and the last three are mean bias-corrected ones. The sigma square and phi are on the log scale.

From figure (2.4) this is evident that the bias of all the parameter has been reduced significantly. Now both types of parameters are well captured. This is an improvement over existing methods where the underestimation problem left unsolved by other aforementioned authors. This is to be noticed that the simple average of the block estimates, $\hat{\theta}_{\texttt{bar}}$ performs almost similar ways as the $\hat{\theta}_{\texttt{fem}}$ and $\hat{\theta}_{\texttt{rem}}$ did. According to our expectations, this should no be the case. The issue is not why the simple average performs equally with other methods that utilize more information, rather the issue is why the methods that exploit more information do not outperform. We flag this issue for further investigation. This could be because the ϕ parameter is chosen small for the experiments. That is the correlation vanishes inside the blocks. However, in the case of

simple average, the spatial dependence between blocks, block sizes, and distribution of the locations inside the block are completely ignored. Although, all the three estimators `bar`, `fem`, and `rem` provide similar values which can be used for prediction. There is another question regarding the estimators which is *robustness*. That is if the estimators perform similarly in different settings. That is, for the different number of blocks and sizes, for low and high parameter values, etc.

In the next chapter we perform some more simulation experiments for assessing spatial asymptotic of `farmer` approach empirically. Also, the outcome of comparative experiments is presented. The results from the real applications are also presented in the next chapter.

Chapter 3

Performance evaluation and real applications

In this chapter, we present the empirical results of a set of simulation experiments, two real applications and comparative performance of `farmer` approach with two existing similar methods. The geostatistical regression model formulated in (2.26) is the target model. Our target is to estimate the parameters of the model with analytical standard errors. All the experiments and applications in the case of Gaussian random fields are considered here in this chapter. The simulation experiments are designed to understand how the `farmer` approach performs in the presence of nugget effect, what if the block size m increases with the fixed number of blocks K, what is the trade-off between increasing m and K. Measuring the time required in `farmer` approach is an objective too. For conducting simulation experiments we have considered the model (2.26) which we recall here for the convenience of readers.

$$Y(s) = X(s)^\top \beta + S(s) + e(s) \tag{3.1}$$

As we know that exact inference in the spatial context is often not available (Cressie (1992)). The alternative is to draw inference based on *asymptotic* results. In the spatial context, there are two types of asymptotic scenarios through which the asymptotic behavior of estimators is studied usually. We introduce them briefly in the next section prior to move to the simulation experiments.

3.1 Spatial asymptotics

Asymptotic results in statistics are derived by letting the sample size n tend to infinity. This can be done in various ways based on situations. When replication involve in the analysis the number of replications is allowed to tend to infinity. In time series this can be done by observing the series for an infinite period. In the spatial setting, this can be done in two ways.

Let us assume, $\mathcal{D} \subset \mathbb{R}^d$ is the domain of observations. In the first case, we can let $n \longrightarrow \infty$ by allowing $|\mathcal{D}| \longrightarrow \infty$, where $|\mathcal{D}|$ is the size of the domain. In this scenario, the number of observation increased by increasing the domain infinitely. This is similar to observing a time series over an infinite period. The number of sample locations per unit area remains finite over space. This asymptotic is called *domain-increasing asymptotics*. Most often in the lattice scheme, this is prohibitive to apply where space is bounded. Such as the data over a specific country where the boundary of the country can not be expanded.

On the other hand, when $0 < |\mathcal{D}| < \infty$, that is the locations are distributed over a finite space $\mathcal{D} \subset \mathbb{R}^d$. In this scenario, we can let $n \longrightarrow \infty$ by infinitely sampling locations between existing locations. In mining, this is known as infill sampling. Therefore, the asymptotic behavior when $0 < |\mathcal{D}| < \infty$ but $n \longrightarrow \infty$ is called the *infill asymptotics*. The minimum distance between points tends to zero as $n \longrightarrow \infty$ in this type of asymptotic.

In case of `farmer` estimators,

- *Infill asymptotic* is represented as $m \longrightarrow \infty$ when K fixed, $0 < |\mathcal{D}| < \infty$.
- *Increasing-domain asymptotic* is represented as $K \longrightarrow \infty \quad$ for fixed $\quad m$.

where, m is the block size and K is the number of blocks.

3.2 Simulation experiments

3.2.1 (a) Checking performance under domain infilling

We have designed this experiment to understand how the `farmer` approach performs in the presence of nugget effect and what happens when the block size m gets larger and larger with the fixed number of blocks K and $0 < |\mathcal{D}| < \infty$. This allows us to check the asymptotic behavior of farmer estimators under infilling. We consider the model (3.1) for simulation experiments. We have fixed the number of blocks to $K = 80$

and repeated the experiments for average block sizes $m \approx \{150, 250, 400\}$. The total number of locations are then $n = \{12000, 20000, 32000\}$ for three different scenarios over the domain $(0, 30) \times (0, 30)$. The size of the block is not the same for all the blocks after splitting the data rather there is a slight variation. The number of blocks 80 is reasonable here. The reason is that we have formulated the $MCAR(\nu, \Lambda)$ model such a way that we required to estimate only one parameter ν whatever be the number of parameters in the original model. We have included a single explanatory variable in the linear spatial regression model which is known and generated from $N(0, 0.5^2)$. The true parameter values are considered as $\beta_0 = \beta_1 = \sigma^2 = 1,\ \phi = 0.15,\ \tau^2 = 0.1$ for all the scenarios. We have implemented the algorithm with spatial dependence parameters $\sigma^2, \phi,$ and τ^2 in log scale which ensures the the estimate can take values over $(-\infty, +\infty)$. Moreover, log transform estimates of these parameters are more probable to closer normal approximation which validates our assumptions for MCAR models. The experiments are repeated 200 times for every case.

The results from the set of experiments are presented in figures (3.1, 3.2, 3.3, 3.4, 3.5) and table (3.1). In the table, we have presented the ten percent trimmed mean. All the versions of estimates for β parameter, perform excellently in all the settings. The variance is reduced with larger block size which is expected. However, performance in the case of spatial dependence parameters is not as good as the regression parameters. This is the usual case in the estimation problem of the spatial model. The dependence parameters are always harder to estimate. We observe that correction of bias is working in the presence of nugget effect however not as faster as before. As the block size increases performance improved. Similar, behavior also observed by Liang *et al.* (2013) and Barbian and Assunção (2017) for their methods. In our case, the improvement is two folds, first bias reduction due to increase block size and secondly employing the bias reduction method. Therefore, with block size $m = 400$ we have reached a reasonable level of bias reduction in both the σ^2 and ϕ. However, for the parameter nugget variance τ^2 improvement is not as faster as the others do. Even with block size 400 the downward bias is still there. Though bias is not reduced that much for nugget effect the variance does. There is one interesting observation for nugget variance that `farmer fem` and `rem` performing better than simple average which is not the case with other parameters. The ratio of ϕ/σ^2 has also been plotted. The performance improves as the block size increases. The difference between bias-corrected and uncorrected decreases with the larger blocks.

This is clear that as the block size gets larger the variance of all the `farmer` estimators including both bias-corrected and uncorrected decreases. At block size, $m = 400$ all

the versions of all the parameters performs better in terms of empirical consistency. Therefore, a flavor of infill asymptotic behavior of the `farmer` estimators is suggestive. That is, as more sample is drawn over a fixed domain the estimate become consistent at least empirically. However, without mathematical derivation, we can not confirm this claim.

Scenarios	Pars(truth)	$\hat{\theta}_{\texttt{bar}}$	$\hat{\theta}_{\texttt{fem}}$	$\hat{\theta}_{\texttt{rem}}$	$\hat{\theta}^{\dagger}_{\texttt{bar}}$	$\hat{\theta}^{\dagger}_{\texttt{fem}}$	$\hat{\theta}^{\dagger}_{\texttt{rem}}$
$K = 80, m = 150$	$\beta_0(1.00)$	0.9994	0.9994	0.9994	0.9992	0.9992	0.9992
	$\beta_1(1.00)$	1.0004	1.0002	1.0002	1.0004	1.0003	1.0003
	$\log(\sigma^2)(0.00)$	-0.0492	-0.0492	-0.0493	-0.0358	-0.0368	-0.0368
	$\log(\phi)(-1.897)$	-1.9182	-1.9182	-1.9183	-1.8554	-1.8548	-1.8549
	$\log(\tau^2)(-2.303)$	-9.1877	-4.0033	-4.0160	-8.4382	-4.0923	-4.0922
$K = 80, m = 250$	$\beta_0(1.00)$	1.0002	1.0002	1.0002	1.0002	1.0002	1.0002
	$\beta_1(1.00)$	1.0001	1.0001	1.0001	1.0001	1.0000	1.0000
	$\log(\sigma^2)(0.00)$	-0.0251	-0.0262	-0.0262	-0.0155	-0.0171	-0.0171
	$\log(\phi)(-1.897)$	-1.9255	-1.9239	-1.9239	-1.8888	-1.8865	-1.8865
	$\log(\tau^2)(-2.303)$	-6.5994	-5.6011	-5.6011	-6.1164	-5.2542	-5.2542
$K = 80, m = 400$	$\beta_0(1.00)$	1.0013	1.0012	1.0012	1.0013	1.0012	1.0012
	$\beta_1(1.00)$	1.0002	1.0003	1.0003	1.0002	1.0003	1.0003
	$\log(\sigma^2)(0.00)$	-0.0161	-0.0183	-0.0183	-0.0071	-0.0098	-0.0098
	$\log(\phi)(-1.897)$	-1.9278	-1.9242	-1.9242	-1.9011	-1.8969	-1.8969
	$\log(\tau^2)(-2.303)$	-4.2135	-3.9946	-3.9946	-3.9728	-3.7902	-3.7902

TABLE 3.1: Estimates obtained from the set of experiments (a)

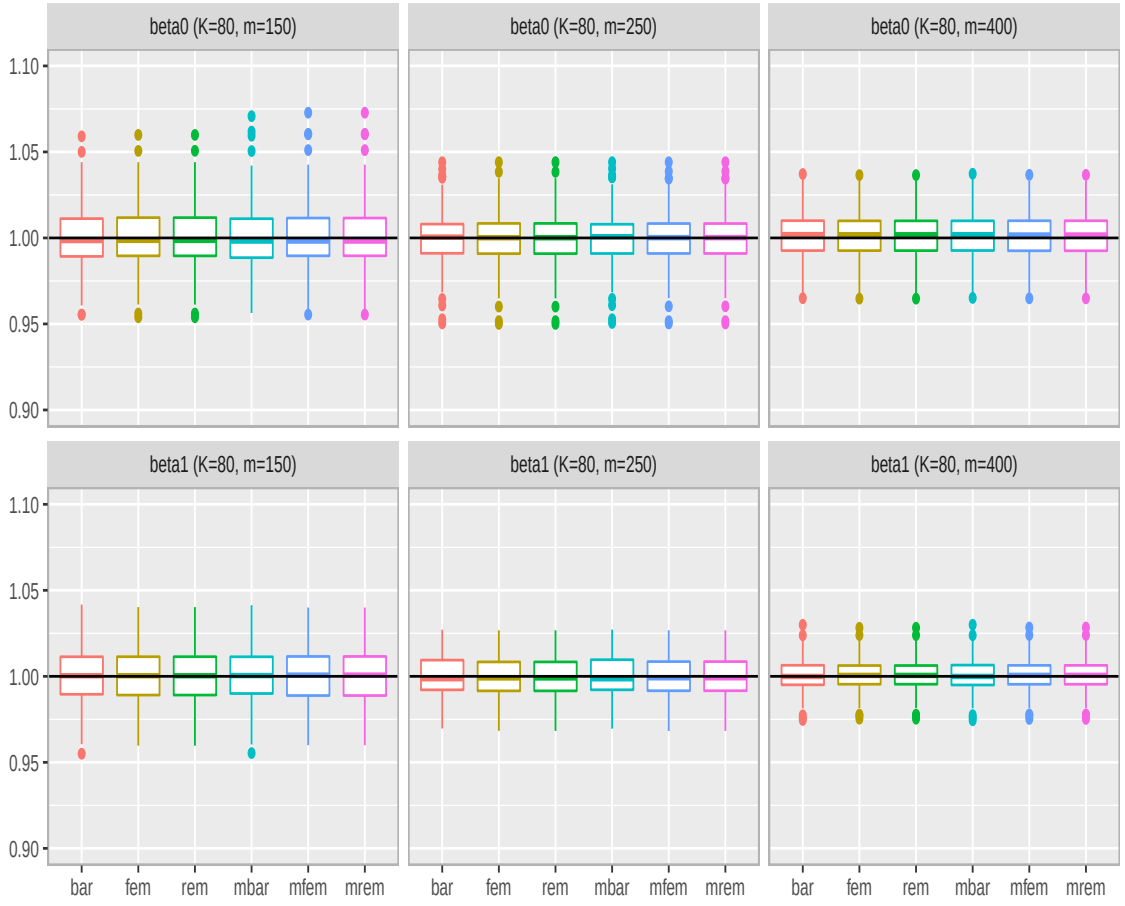

FIGURE 3.1: Box-plots represent the outcome of experiment (a). Each of boxes represents different versions of estimates of (β_0, β_1).

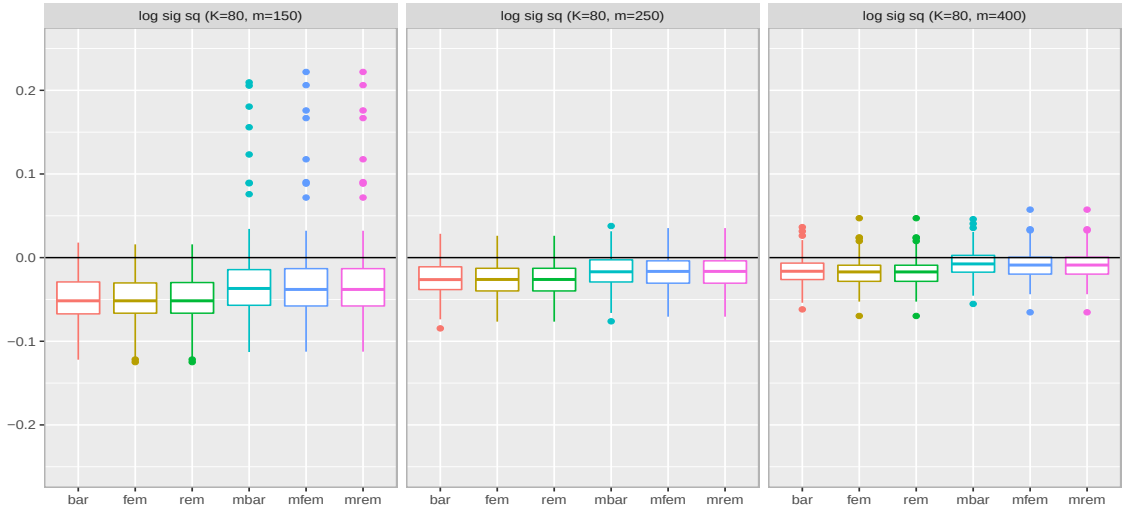

FIGURE 3.2: Box-plots represent the outcome of experiment (a). Each of boxes represents different versions of estimates of $\log(\sigma^2)$.

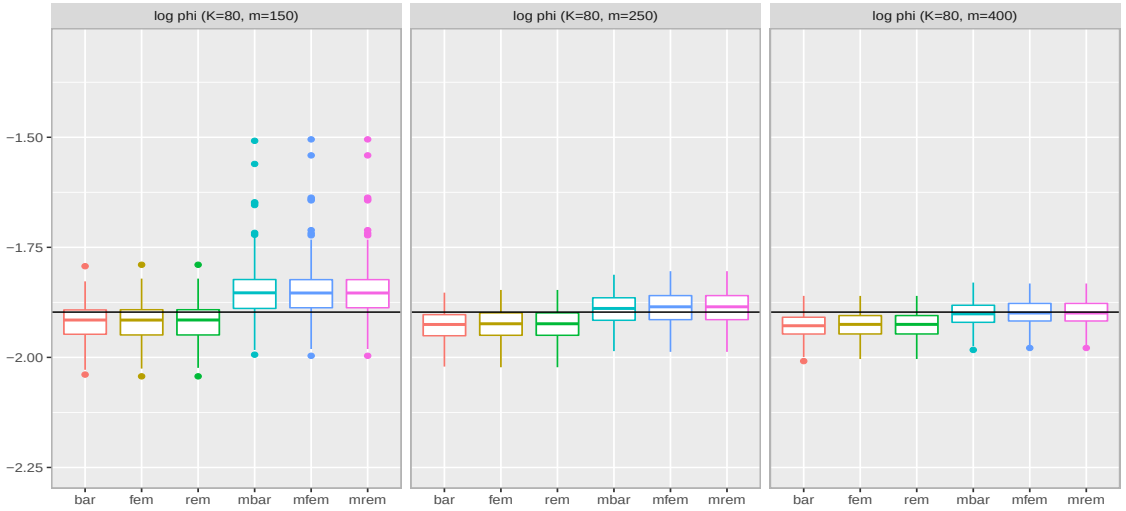

FIGURE 3.3: Box-plots represent the outcome of experiment (a). Each of boxes represents different versions of estimates of $\log(\phi)$.

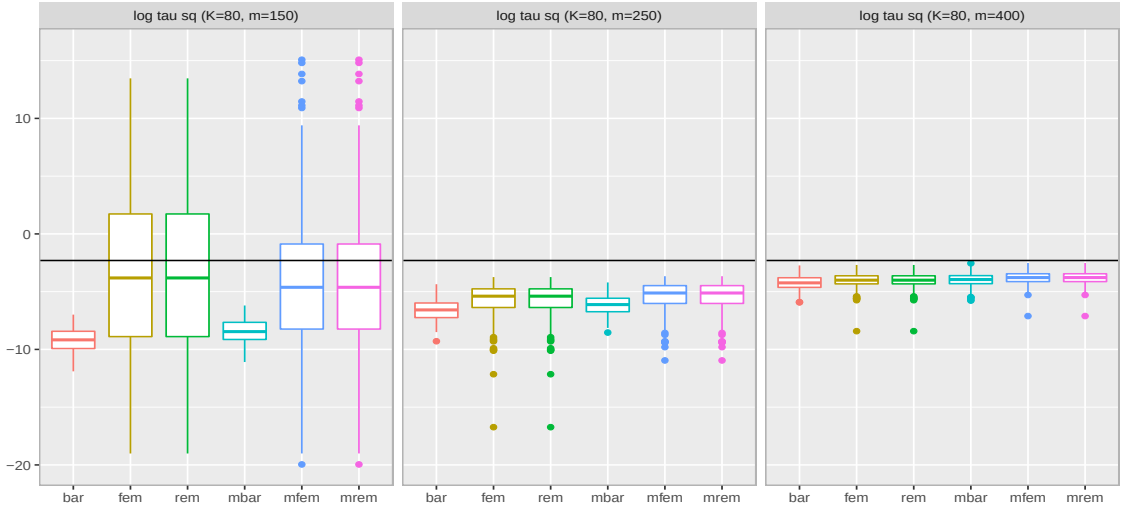

FIGURE 3.4: Box-plots represent the outcome of experiment (a). Each of boxes represents different versions of estimates of $\log(\tau^2)$.

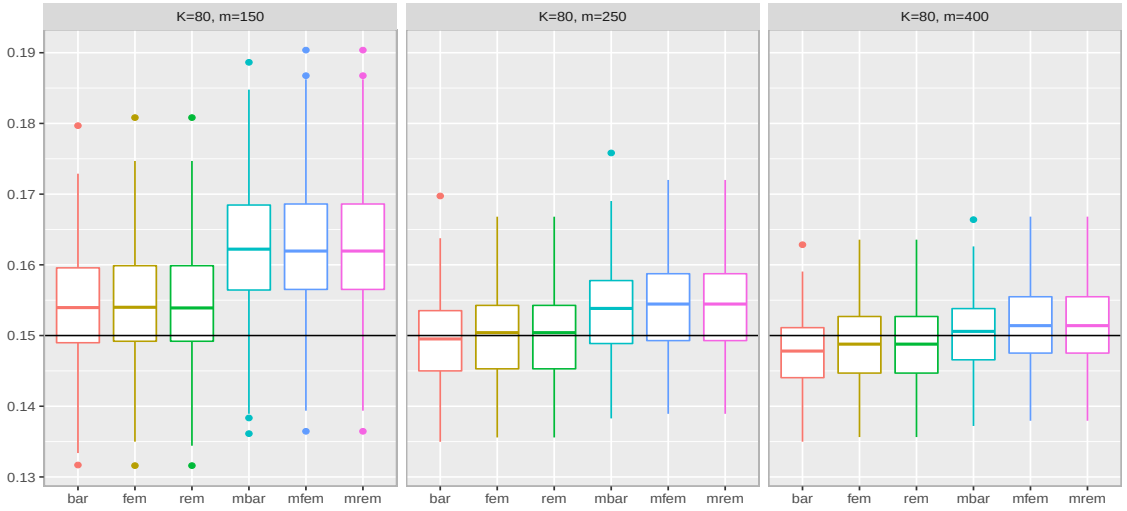

FIGURE 3.5: Box-plots represent the outcome of experiment (a). Each of boxes represents different versions of estimates of ϕ/σ^2.

In summary, we observed that,

- Downward bias has been reduced by a reasonable amount however not removed completely. This was a known problem in literature and through the proposed approach solved partially.
- Estimates are found to be consistent at least empirically under infill asymptotic. To be confirmed, the mathematical proof is required.
- The nugget parameter is not improved that much in terms of bias reduction.

Now, we present another set of experiments in the next subsection for assessing the performance of `farmer` estimators under increasing domain settings.

3.2.2 (b) Checking performance under increasing domain

In this set of experiments, we have considered the same model (3.1) and specifications in terms of true values and number of experiments as done in (a). However, we have fixed the block size m equal to 250 and conducted the experiments for different values of $K = \{48, 300\}$. These leave us the $n = \{12000, 75000\}$. The results are presented similar way using box plots. To see the patterns we have presented the box plots for $K = \{48, 80, 300\}$, where results for $K = 80$ are taken from the experiments set (a). For the first scenario, we have generated Gaussian random field over the domain $(0, 18)\times(0, 18)$ and for the last scenario, we have expanded the domain to $(0, 60)\times(0, 60)$. For $K = 80$ we have domain $(0, 30) \times (0, 30)$ in the previous experiments.

The performance in the case of regression parameters is similar to before. However, in the case of spatial dependence parameters behavior contrasting to the previous experiments is observed. The variance parameter σ^2 is overestimated when the number of blocks gets larger. On the other hand, the scale parameter ϕ is more underestimated in the same situation. One thing is common that the variance of the estimates reduced with larger number of blocks which again provide empirical evidence of consistency under increasing domain setting. The increasing number of blocks has added some extra bias to all the spatial dependence parameters in different directions. The added bias is positive to the variance parameter however negative to the scale and nugget parameters.

In both the experience set (a) and in (b), the value of the parameter ϕ is chosen a bit smaller. However, when we have compared the performance in the next section a bit larger value ($\sigma^2 = 3.0, \phi = 0.2$) is chosen. Also, we have performed some more simulation considering $\phi = 2.0$ which is quite large. The results of these simulation experiments are presented in . In that simulation, we found similar performance for regression parameters however for spatial dependence parameters the results are not as convincing as before. There is downward bias in all three dependence parameters. For σ^2 and ϕ, we obtain the estimates with some positive bias. The nugget parameter τ^2 and the ration ϕ/σ^2 performs better with bias correction.

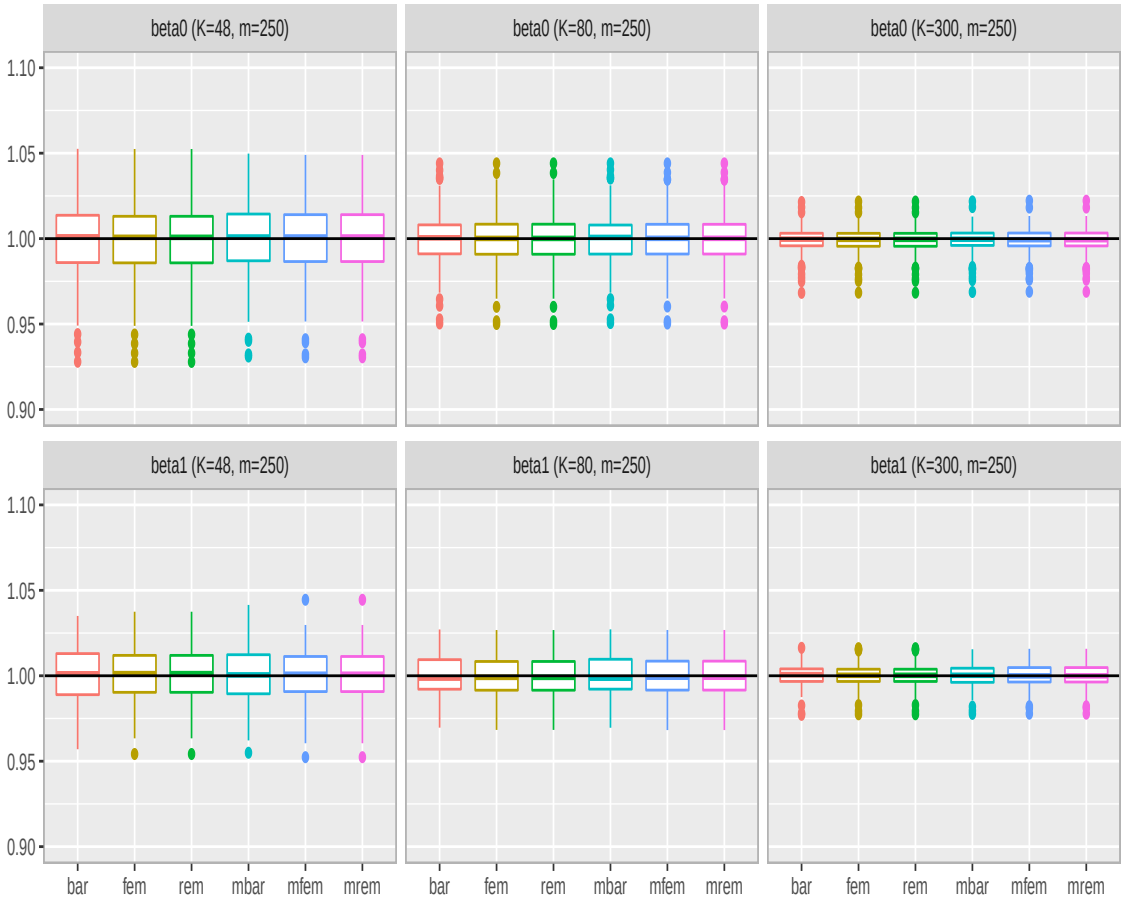

FIGURE 3.6: Box-plots represent the outcome of experiment (b). Each of boxes represents different versions of estimates of (β_0, β_1).

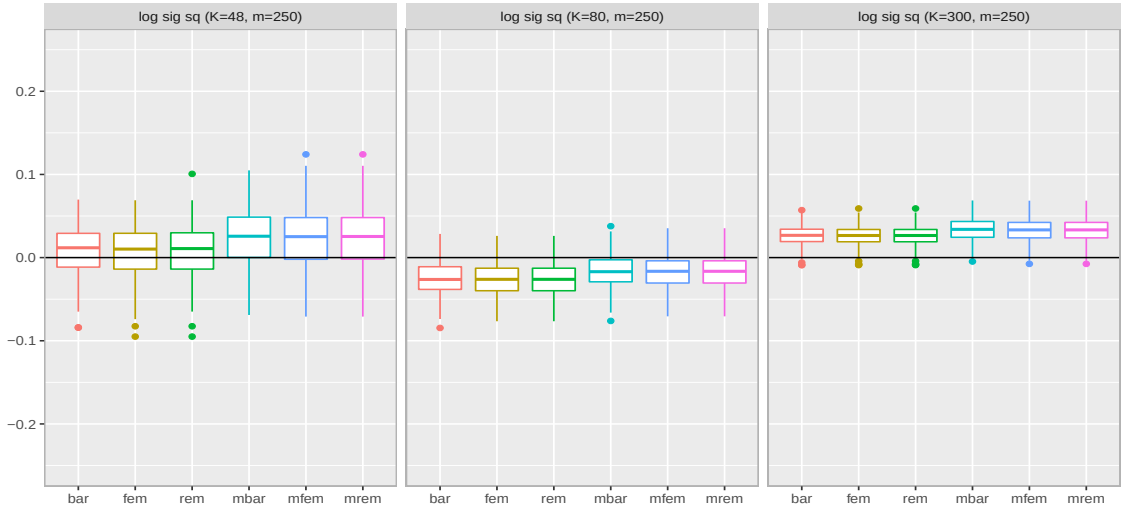

FIGURE 3.7: Box-plots represent the outcome of experiment (b). Each of boxes represents different versions of estimates of $\log(\sigma^2)$.

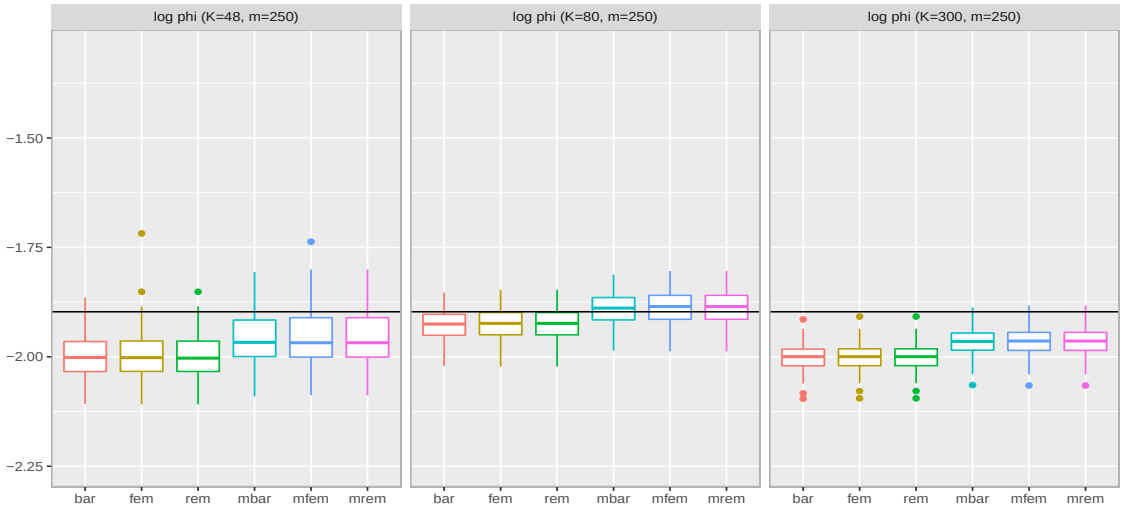

FIGURE 3.8: Box-plots represent the outcome of experiment (b). Each of boxes represents different versions of estimates of $\log(\phi)$.

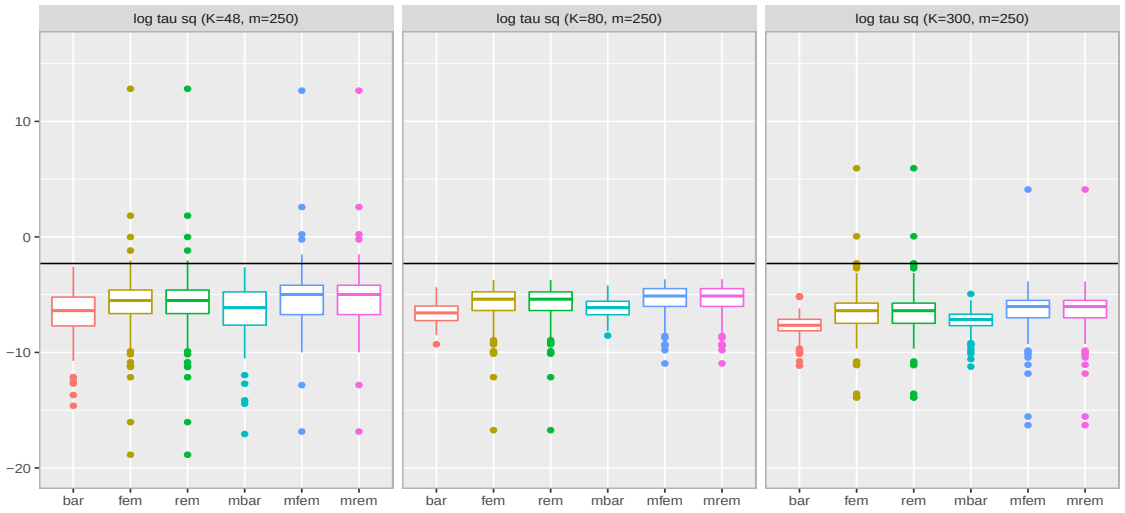

FIGURE 3.9: Box-plots represent the outcome of experiment (b). Each of boxes represents different versions of estimates of $\log(\tau^2)$.

The possible interpretation of this results could be due to negative correlation between σ^2 and ϕ.

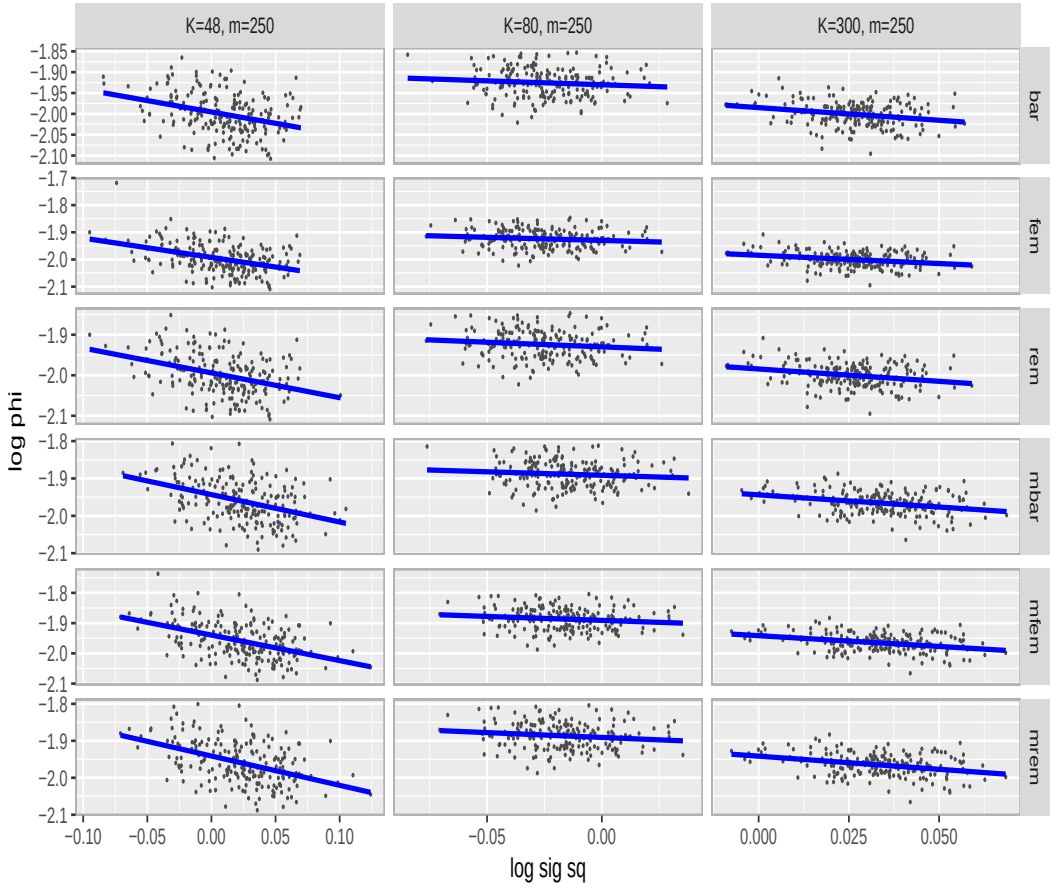

FIGURE 3.10: Scatter plot with smoothed line between $\log(\sigma^2)$ and $\log(\phi)$, outcome of experiment (b).

The association is negative which is also observed from the box plots above. Therefore, this graph suggests that the positive bias in the variance induce negative bias in scale or vice versa.

Based on the experiments sets in (b), the following points are observed,

- The regression parameters are empirically unbiased and consistent under the increasing-domain setting.
- The estimates for the spatial dependence parameters are empirically consistent however extra bias has been added at increasing domain scenario.
- Another note is that the `farmer` approach works for the lesser number of blocks as well.

In the next section, we discuss the variance estimation procedure for the estimate. If the working model for meta analysis models coincides with the assumed true models then this estimation is not required by the approach.

3.2.3 Variance estimation and internal efficiency of `farmer` estimators

In the thesis, our focus is on both reducing computational burden and estimating more realistic standard errors of the estimate. Realistic standard error estimation in the subsampling scheme is lacking. Liang *et al.* (2013) has provided no specific guideline for estimating standard error for their estimate rather suggested to repeat the estimating process many times and compute the standard error empirically. On the other hand, Barbian and Assunção (2017) proposed an approach based on the covariance matrix computed from subsamples. There is similarity between the variance estimation in the latter approach and the variance estimation for `farmer fem` estimator though the weights are completely different. We have estimated the standard error for `farmer` estimators from simulation experiments of model (3.1) with $n = 20,000, K = 80, m \approx 250$ and true parameter were considered same as before ($\beta_0 = \beta_1 = \sigma^2 = 1, \phi = 0.15, \tau^2 = 0.1$). We have applied the SHAC estimation procedure for estimating the variance of the estimates as proposed in the second chapter. Considering the simplicity we have applied the Tukey-Hanning kernel with distance matrix as the distance among the centroid of the blocks. We have computed the variances considering two scenarios for the distance scale factor $d_n = \{5, 10\}$. We have presented the 95% confidence interval of `farmer bar, fem`, and `rem` in figure (3.11). Both the analytical and simplified versions are presented.

FIGURE 3.11: `farmer` estimators with respective 95% CI are plotted. From top to bottom CI for $\beta_0, \beta_1, \log\sigma^2, \log\phi, \log\tau^2$ are plotted respectively. Among two main columns, in left 95% CI of bias uncorrected and their simplified versions with $d_n = 5$ are presented and in right same with $d_n = 10$ are presented. The CI for `bar, fem, rem` are presented from left to right in each of blocks. In the $x-$ axis the iteration numbers are presented.

From the figure, we notice that all the `farmer` estimators provide similar results for both the regression as well as the spatial dependence parameters except the nugget parameter. In case of nugget parameter the `fem` and `rem` produce wider CI than that of `bar`. However, we can not say based on this that which one is closer to true. This is suggestive that any form of the `farmer` estimators can be applied when the locations are uniformly distributed over the space. Also, there is no visual difference between the CI estimates for different values of the distance scale factor. This is empirical evidence that the standard error of `farmer` estimators are robust to the choice of d_n which is not expected though. This could be because the variance is driven by the block fisher information matrix. Therefore, the kernel in the SHAC estimator does not have that much influence on this experiment. The simplified variances are found similar to those of analytical ones. This is supportive of assumptions in equations (2.2) and (2.8), that is assumed model is somehow similar to the true model. Therefore, it is reasonable to assume the normality of the block estimates.

Till now all the simulation is done over location uniformly distributed. Simulating over non-uniformly distributed locations could give more insight. This can be done using the locations from a real dataset and assigning the realization of a random variable to each of the locations. We skip this part now and focus on completing other necessary parts of methods comparison and real applications.

We have a good experience that the `farmer` approach is time efficient and provides an analytical solution for estimating the confidence interval. Comparing with other similar methods we would be able to say firmly about the performance of the proposed approach. The next section focus on comparing the performance of `farmer` approach with two other similar approaches.

3.3 Comparison of performance

In this section, we compare the performance of `farmer` approach with two approaches proposed by Liang *et al.* (2013) and Barbian and Assunção (2017). The description of these methods can be found in chapter 1 and the reference thereof. These two are chosen because they are in the same domain of literature as ours. The first one is known as resampling based stochastic approximation (RSA) and the latter one is defined as spatial subsemble (SpSub) estimator. Both of these methods aimed to handle large spatial data. The main strategy behind these methods is subsampling however in different ways. The RSA iteratively updated the estimator based on a new subsample. On the other hand, SpSub combines the subsample outcome using a weighted average scheme. We have

conducted simulation experiments for comparison purposes.

Again, we have chosen the geostatistical model (3.1) for simulation study. The true values of the parameters are chosen as, $\beta_0 = \beta_1 = 1$ which is similar as before. The single known explanatory variable are the realization from $N(0, 0.5^2)$ is considered as done before. We have considered comparetively larger values of variance and scale parameters as, $\sigma^2 = 3, \phi = 0.2$ and the $\tau^2 = 0.1$ remains same. We have generated Gaussian random fields at $20,000$ irregular uniformly distributed locations over the domain $(0, 30) \times (0, 30)$.

For implementing the `farmer` approach we have split the domain into $K = 80$ mutually exclusive and exhaustive blocks which leave us $m \approx 250$ block size. The spatial dependence parameters are estimated on the log scale. The RSA is implemented through the archived `R` package `RSAgeo` version 1.2. To keep similarity, we choose the subsample size is equal to 250 which is a bit lower for this method. For estimating the parameters we have run 2500 iterations with warm up parameter equal 20 and the stepscale parameter is 40 which is following the original article by the authors. For spatial subsemble implementation the authors Barbian and Assunção (2017) have kindly shared their `R-code`. Among the various subsampling techniques, we have chosen five centers technique for comparison purpose. There is no issue of comparing just one splitting technique because they all are comparable among themselves. Similarly, we have selected the subsample of size 250 for this method. For all three methods, we have repeated the experiments 100 times. The computation is conducted in a personal computer with Ubuntu 18.04 operating system, 2.7GHz core i7 processor. The estimates are presented using the box plot below.

From the figure (3.12) we observed that the `farmer` method performs similarly with larger variance and scale parameter. This is an indication of the robustness of the method. The computation time is also within 2 minutes limit. Both the regression parameters and spatial dependence parameters captured very well however downward bias is still there in nugget variance. The `farmer` estimators are empirically consistent as well.

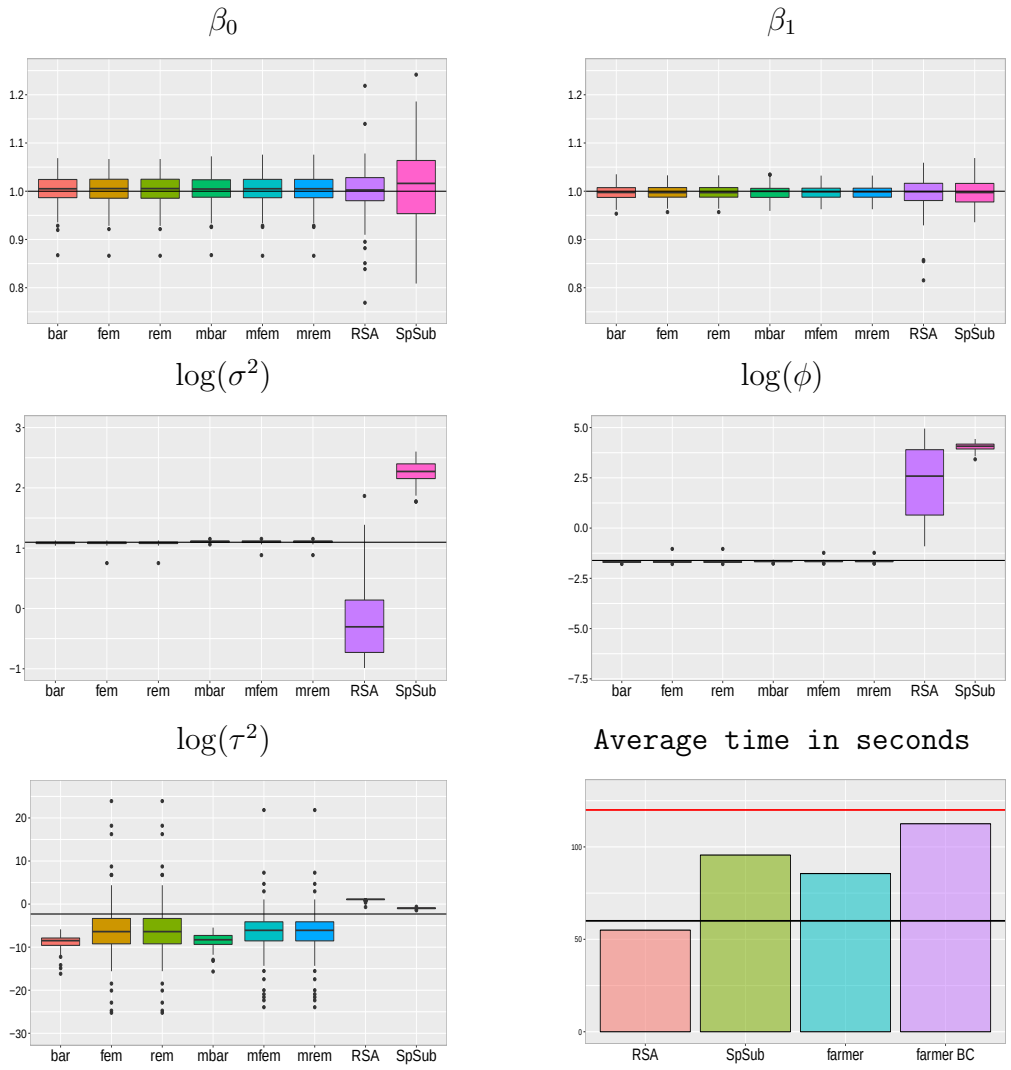

FIGURE 3.12: Comparison of `farmer` method with RSA method by Liang *et al.* (2013) and SpSub method by Barbian and Assunção (2017). The bar plot in the bottom right corner represents the average time in seconds required for computation in each run for various methods.

The graph shows the superior performance of the `farmer` approach in terms of biasness, consistency, and efficiency. The regression parameters are estimated well by every method without any bias however the length of the box suggests that the RSA and SpSub produce variances larger than that of `farmer` approach.

In estimating the spatial parameters `farmer` approach is more consistent and efficient than RSA and SpSub approaches. The `farmer` approach estimates the σ^2 and ϕ with minimal bias and variance while other methods produce large bias and large variance as well. This could be improved with larger subsample size. The authors reported their original articles about this biasness. They also reported that with larger subsample size the bias reduced and produce less variance. This is evident that the `farmer` approach is

faster in reducing bias and variance. Regarding the nugget parameter τ^2, all methods suffer from downward bias. Even after bias correction applied to the `farmer` approach, bias is still there. One possibility to try the median bias correction approach (Kosmidis *et al.* (2017)) at block level with `farmer` approach could improve the results.

The average time in seconds required to complete a single experiment is presented by methods in bar chart. The black and red horizontal lines represent 1 minute and 2 minutes limit respectively. The chart shows that the highest time required for `farmer` bias-corrected approach however it is below the red line. The RSA requires less than 1 minutes even however with a single experiment this method does not provide the estimate of standard error. To get the estimate of standard error it is required to run the method a reasonable number of times and calculate the empirical standard error. In that case, the required time to be multiplied by the number of runs.

Based on the experiments sets in (a), (b) and comparative study following points are observed,

- The regression parameters are easy to estimate with `farmer` approach. They are unbiased and consistent at least empirically under both the infill and increasing-domain asymptotics.
- The `farmer` approach provides bias reduced and empirically consistent estimates for the spatial dependence parameters under infill asymptotic. On the other hand, the method provides empirically consistent but biased estimates under increasing-domain scenarios.
- The computational burden has been reduced by a great amount. The computational gain is also achieved by other methods.

In the next section we present two real data applications of `farmer` approach. We also apply the RSA (Liang *et al.* (2013)) for comparison purposes. The results are compared with MLE obtained from the entire dataset

3.4 Real examples

Two examples covering two diverse fields, climate and health are considered. In the first example, we have fitted the Gaussian geostatistical model to the precipitation data and in the second model, we transformed the river-blindness case count data to logit and fitted Gaussian geostatistical model to that transformed data. The details of the process are described in the following two subsections.

3.4.1 US precipitation data analysis

The same data set has also been used by several authors for different purposes (see for example Liang *et al.* (2013), Johns *et al.* (2003), Furrer *et al.* (2006), Kaufman *et al.* (2008)). A cleaned version of the data is also available in `fields` R-package titled `USprecip`. The publicly available (see `www.image.ucar.edu/GSP/Data/US.monthly.met/`) data set contains raw monthly total precipitation measured in millimeters and precipitation anomaly for April 1948 in 11,918 locations over the contiguous United States. The longitude and latitude are also available with the data set. The precipitation anomaly is defined as the monthly totals standardized by the long-run mean and standard deviation for each station. The locations of the data are as follows.

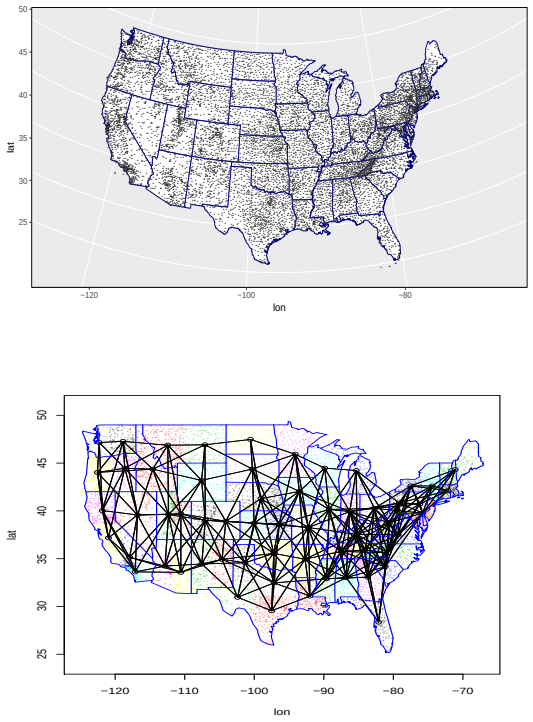

FIGURE 3.13: US precipitation data location and generated blocks

We have excluded 190 locations which are observed outside the national boundaries. This decision is taken because we have created the blocks considering the natural boundaries. Using `map.where()` function we could not identify these locations associated with any state. Finally, we have analyzed the 11,728 locations and have split the region into 42 block. This number comes considering the minimum block size of 200 observation however at the end we have reached an average block size of $m \approx 280$. To split the region into blocks we keep the natural boundaries unbroken rather the states with the larger number of locations are split into several blocks. The states of size 200 *to* < 400

have been considered as a single block. The states of size ≥ 400 split up into blocks of average size ≥ 200. In this process of splitting, we obtained smaller blocks. The states with size lower than 200 are merged and formed blocks of size greater than 200. We have constructed the neighborhood structure of the blocks considering that a block should be associated with surrounding blocks. To ensure this we have considered a distance within which if the centers of any neighbor blocks fall we considered they are connected. The distance between two points is calculated based on longitude and latitude and 8 is considered as the criteria within which if two centers fall they are considered as associated. The number 8 is chosen based on the fact that every neighbor of a block should be connected. Based on this we have constructed the adjacency matrix for the second level of modeling, that is the $MCAR$ model.

We have considered the following model keeping similarity with Liang *et al.* (2013),

$$Y(s) = \beta_0 + S(s) + e(s), \tag{3.2}$$

where $S(s)$ Gaussian process which follows exponential covariance model where σ^2, ϕ are the variance and scale parameters respectively. The nugget variance τ^2 is considered here and estimated.

We have implemented the `farmer` approach through algorithm (1). At the same time the RSA method is also applied using `RSAgeo` package in `R` with subsample size 280, 2500 iterations, stepscale parameter 40 and warmup parameter 20. The results are presented in table (3.2) and figure (3.13). The MLE on entire data set were obtained by Barbian and Assunção (2017) for the same data set however on randomly selected $11,000$ locations. We have included MLE as well from that article directly due to computational issue.

BC	ests	β_0	$\log(\sigma^2)$	$\log(\phi)$	$\log(\tau^2)$	time(s)
	farmer bar	0.0615	-0.8051	0.4031	-3.8824	
	SE	0.0925	0.0532	0.0601	0.2257	
no	farmer fem	0.0816	-0.8213	0.4516	-3.6360	38.15
	SE	0.0801	0.0434	0.0297	0.2216	
	farmer rem	0.1055	-0.8332	0.4503	-3.5841	
	SE	0.0695	0.0404	0.0260	0.2259	
	farmer bar	0.1622	0.2931	1.5336	-3.8451	
	SE	0.0868	0.3837	0.4032	0.2175	
yes	farmer fem	0.1654	0.0944	1.3922	-3.6187	46.84
	SE	0.0739	0.2821	0.2868	0.2181	
	farmer rem	0.1946	0.0432	1.3523	-3.5671	
	SE	0.0695	0.0404	0.0260	0.2259	
	RSA	0.1570	-0.1865	1.1663	-2.6683	2327.97
	SE	0.0170	0.0401	0.0803	0.0788	
	MLE	0.256	1.0872	2.3025	-3.0159	1.32e6

TABLE 3.2: Estimates of parameters and their standard errors obtained using `farmer` approach ($K = 42, m \approx 280$), RSA ($m = 280$) and MLE. Estimation times in seconds are also reported in the last column.

From, the table this is clear that there is variation among methods and even within methods for different scenarios. Also, no method captures the MLE rather they are underestimated. The results from this application seem to are not consistent with simulation experiences. From figure (3.14) we notice that the confidence interval for bias-corrected `bar` is even wider for some parameters. Unfortunately, the confidence interval for RSA is too narrow that it does not cover the MLE for any parameter. This could be due to the reason that the same algorithm is run multiple times on the same data and the empirical standard error calculated. Barbian and Assunção (2017) also experienced lots of variation for different subsample size as reported in their article. This is still a mystery. However, a note about this dataset is that around 50% of the data are infilled using some mathematical model. This may explain some extents of the cause of variations of findings.

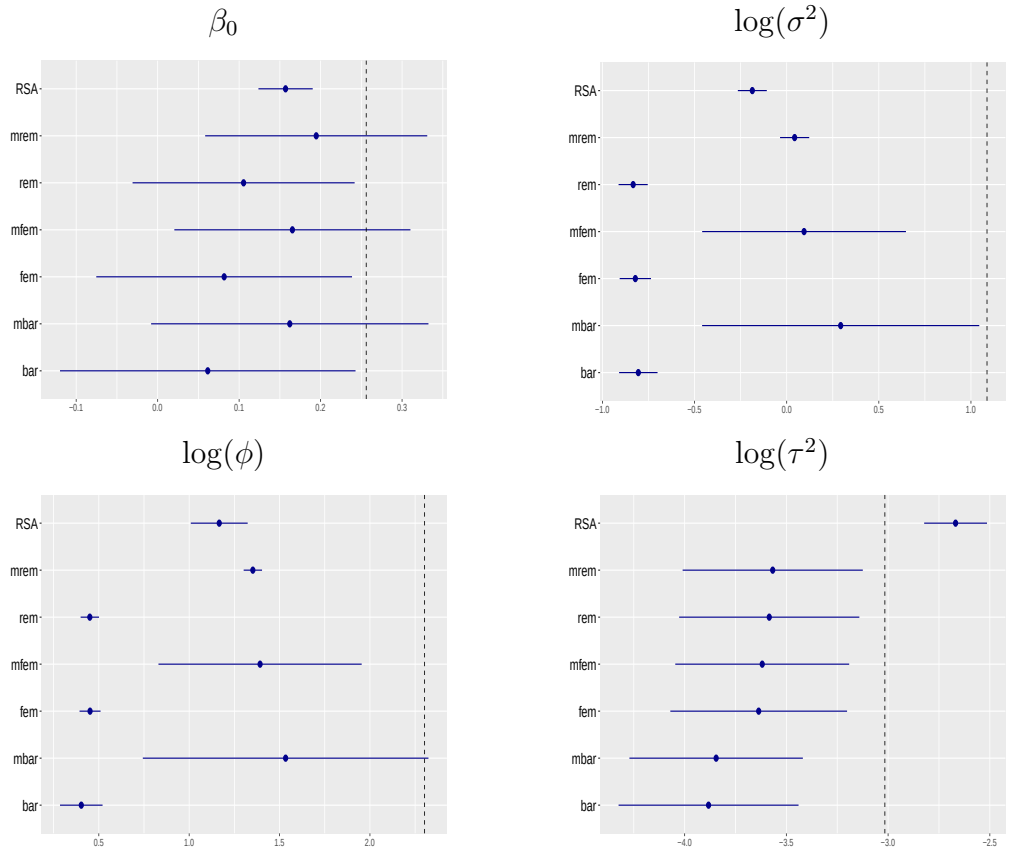

FIGURE 3.14: US precipitation data: comparison of various estimates with MLE(dashed line).

There is an interesting pattern in the above figure. Almost all the bias-corrected `farmer` estimators capture or close to capture the MLE. In the next subsection, we present another real application of `farmer` approach to the Gaussian geostatistical model.

3.4.2 *Onchocerciasis* data over 18 African countries

Onchocerciasis is a disease caused by *Onchocerca volvulus*, which is a worm (filaria). Usually, the human eye and skin are affected by this worm. *Simulium species* which is the scientific name of a blackfly is the main culprit for it which transfers the larvae of the worm into the human body. These flies breed in fast-flowing streams and rivers. Therefore the individuals living nearby are mostly at increasing risk of blindness, hence the name *river*

FIGURE 3.15: Macrofilariae(left), an adult blackfly(right)

blindness. Thousands of larval worm or baby(microfilariae) are produced inside the human body from a single adult worm(macrofilariae) and migrate to the skin and eye. Various eye and skin problems caused by the toxicity of dead microfilariae. Such as terrible itching, lesions, repeated occurrence of which could lead to irreversible blindness. Various dis-figuration of skin such as "leopard" skin and "lizard" skin occurred due to the toxicity of dead microfilariae.

About 50% of men over the age of 40 years in some West African communities, had been blinded by the disease and the people from the affected area had to move in a less productive upland country which induces the economic losses. Onchocerciasis Control Programme in West Africa (OCP) was a successful program for reducing the problem by a large scale in West Africa. The African Programme for Onchocerciasis Control (APOC) was created to implement community-based treatment with *ivermectin* in all remaining areas in Africa where onchocerciasis was a public health problem.

This dataset contains information on $13,681$ villages over 18 African countries. At every village, $30-50$ adult males were sampled and tested for the presence of nodules. This produces two variables `Ex` and `Pos` which represent the number of adults examined and number diagnosed as positive. The geographic coordinates longitude and latitude of the center of the village were captured using Global Positioning System (GPS). The locations of the villages are presented in the figure below. The generated blocks and their networks are shown in chapter 2.

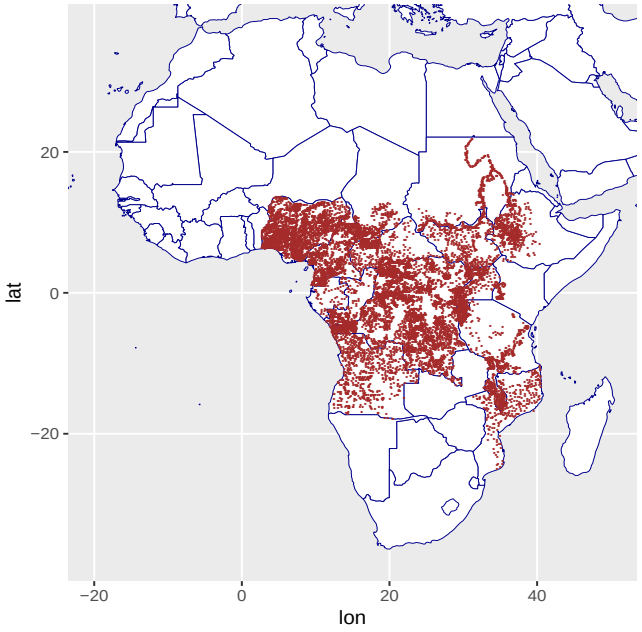

FIGURE 3.16: River blindness data test locations spanned over 18 countries of Africa

The observed and estimated distribution of palpable nodules with other details is

presented in detail in Zouré *et al.* (2014). To show the application of `farmer` approach for Gaussian data we have converted the data to *logit* using the following formula,

$$logit = \log\left(\frac{\texttt{Pos} + 0.5}{\texttt{Ex} - \texttt{Pos} + 0.5}\right). \tag{3.3}$$

In both numerator and denominator, 0.5 is added to avoid the infinite value produced by log function. We followed the similar mechanism of data splitting as done before for precipitation data, that is we split inside the natural boundaries of countries. For this example, we have considered the splitting criteria to 500 locations. That is if the number of test locations for a specific country exceeds the 500 limit we split the country. For example the country *DR Congo* we obtained 20 blocks. Also, there are some countries where the number of sample locations is very less, in that case, we have merged with the nearest country. Such as *EQUATORIAL GUINEA* and *GABON* has 134 and 59 locations therefore we have merged these two countries to create a single block. In this way, we keep the natural boundary unbroken. We found in total 61 blocks in total with average size $m = 235$.

After transforming the data to *logit* we have applied the `farmer` approach, RSA and MLE on entire data set for estimating the geostatistical model (3.2) and the results are presented in the table (3.3) and figure (3.17). The RSA is applied with subsample size $m = 250$, 2500 iteration, stepscale parameter 40 and warmup parameter equal 20. To estimate the standard error of the RSA method we have repeated the method 25 times and the empirical standard error is reported. The MLE is obtained using `PrevMap` package in the cluster. This data was used by Noma *et al.* (2014) and Zouré *et al.* (2014) with a view to Rapid Epidemiological Mapping of Onchocerciasis. More details of the original objectives and data collection procedure can be found in the aforementioned two references.

BC	ests	β_0	$\log(\sigma^2)$	$\log(\phi)$	$\log(\tau^2)$	time(s)
	farmer bar	-2.1145	0.2229	-0.3797	-0.3725	
	SE	0.1121	0.0000	0.0611	0.0525	
no	farmer fem	-2.3717	0.1648	-0.4048	-0.4325	48.044
	SE	0.0760	0.1444	0.1163	0.0818	
	farmer rem	-2.4894	0.0964	-0.4034	-0.4485	
	SE	0.0000	0.1801	0.1299	0.0702	
	farmer mbar	-2.1628	1.2620	0.7566	-0.3565	
	SE	0.1165	0.1900	0.2172	0.0531	
yes	farmer mfem	-2.3445	0.9226	0.5386	-0.3988	56.121
	SE	0.1045	0.2030	0.1713	0.0826	
	farmer mrem	-2.4495	0.8254	0.5263	-0.4135	
	SE	0.0583	0.2092	0.1686	0.0723	
	RSA	-1.9524	1.2605	1.5885	0.2823	1404.416
	SE	0.8116	0.5844	0.9156	0.1144	
	MLE	-2.8027	0.9232	0.2624	-0.1791	22204.87
	SE	0.1568	0.0710	0.0826	0.1437	

TABLE 3.3: Estimates of parameters and their standard errors obtained using `farmer` approach ($K = 61, m \approx 235$), RSA ($m = 250$) and MLE. Estimation times in seconds are also reported in the last column.

In the table (3.3) and figure (3.17) we have presented the β_0 and log of spatial dependence parameters σ^2, ϕ, τ^2 for all methods. The standard error of the log-transformed estimates are also in log scale. The solid vertical line is the MLE and two other dashed vertical lines represent the 95% confidence interval of MLE. The standard errors of `farmer` estimators are reasonably similar to that of MLE however for RSA the standard error is a bit higher. Larger subsample may solve this problem as the author suggested in the original paper (see Liang *et al.* (2013)).

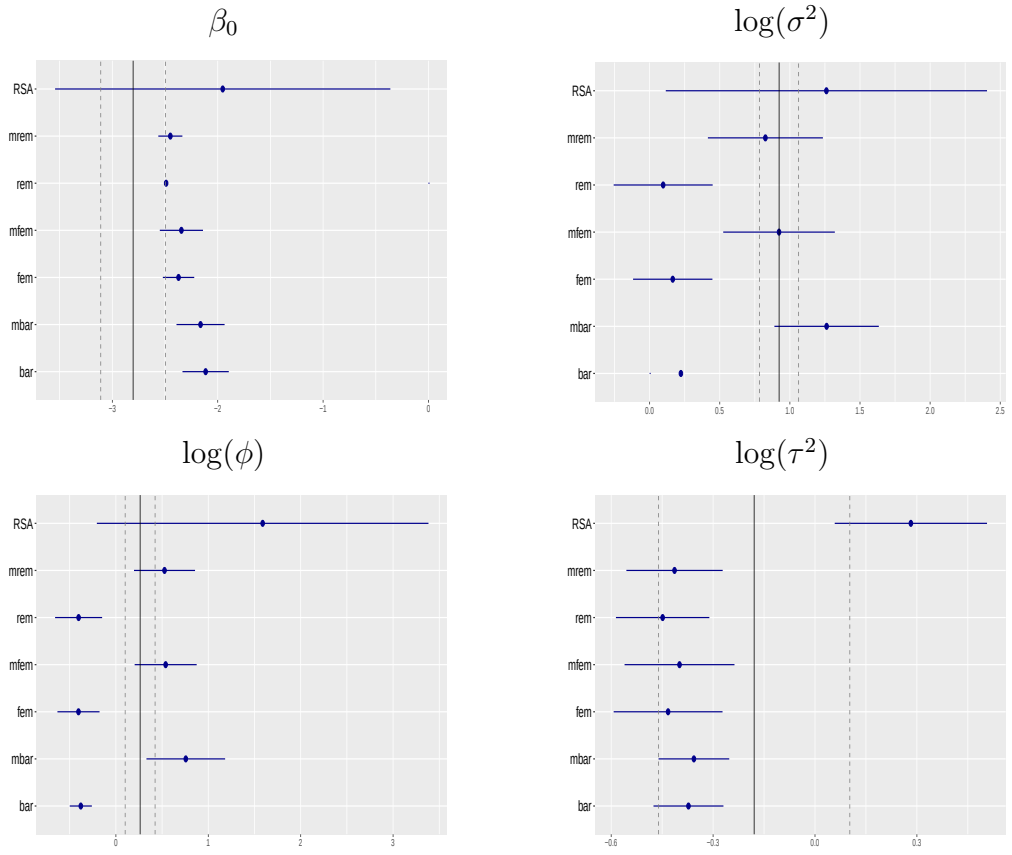

FIGURE 3.17: African river-blindness data: comparison of various estimates with MLE(solid vertical line) and 95% CI of MLE (dashed vertical lines).

The figure suggests that for β_0, `farmer` `fem` and `rem` overlap their confidence interval with MLE however `bar` does not. The RSA is the farthest one from the MLE however the confidence interval overlaps with the interval of MLE though the interval for RSA is much wider. `farmer` method works better in the case of spatial parameters, especially after bias correction. For τ^2 all the estimates are inside the interval of MLE except RSA although the error of MLE is larger for this parameter.

Therefore, based on simulation experience and real application results it is easily recommended to use `farmer` approach for large spatial data, which provide less biased and consistent estimates at least empirically. Also, the approach provides a more realistic standard error of the estimate which is very important for making the inference. Moreover, the method is time efficient. The average time for a moderate to large data set $(n = 20,000)$ is around two minutes.

With this application, we would like to close the chapter and topic on Gaussian `farmer`. In the next chapter we introduce `farmer` approach for non Gaussian data. Non Gaussian is data are more frequent in global health and social studies.

Chapter 4

The `farmer` approach for non-Gaussian data

4.1 Motivation

The `farmer` approach is expanded for non Gaussian data of many types, specially binomial or Poisson type count. Method to deal with the large non-Gaussian data is less frequent however the non-Gaussian data is not less frequent. In spatial setting the authors have not seen any methodology for handling large non-Gaussian spatial data yet however the existing subsampling approaches may be adaptable.

In the real field, there are many problems where the Gaussian model is not appropriate. Rather the variable of interest is binary or count. For example, the presence or absence of a disease, the number of diseased people in an area. These types of data are very common in practice and for analyzing them the classical models are (a) Binary logistic, (b) Binomial logistic, and (c) Poisson models. Based on sampling assumptions and structure of data different versions of these models have been evolved. As long as the independence assumption holds the model estimation, prediction and inference are simple otherwise the situation becomes a bit complicated.

Diggle and Giorgi (2019) in their book dedicated a section on motivating examples at the beginning chapter. In that section, the authors pose a problem of mapping river-blindness data from 20 African countries which is a big data problem indeed. The data is not Gaussian and geographic location is associated with each data point. The river-blindness data has already been presented in chapter three. In public health problems, one of the primary targets is to produce a map to identify the potential hot spots.

We would like to present another motivating ongoing problem that needs to be considered immediately. At this moment while we are writing this section, Bangladesh is

facing one of the biggest dengue outbreaks. Especially the capital is under severe attack. From January to August 2019 around 30,000 people have been hospitalized with the disease over the country and in the past week around 2,000 patients have been admitted to hospitals every day[1]. Recently died more than 100 of this disease. The conditions of hospitals are presented below:

FIGURE 4.1: Snaps of hospital and surroundings in Dhaka, Bangladesh during dengue outbreak, August 2019. The beds are full and patients are on the floor, the huge queue outside the hospital for the diagnostic test for dengue.

The government agencies and various NGOs are trying to control the situation however the situation getting worse gradually. This suggests that efforts should be given in the right way. Analyzing real-time data could help to improve the situation. One solution could be to identify the priority areas based on prevalence mapping and to administer the anti-mosquito medicine.

This is a real big data problem and to handle this huge amount of data needs a computationally efficient approach. When the patients are getting admitted to hospitals they are leaving their demographic information, locations, etc. This huge amount of data can help to find out the solutions.

Diggle and Giorgi (2019) showed that mapping based on the geostatistical model is more useful than the simple map. More useful information can be generated from the model-based maps. When a statistical model enters into the problem, the task of estimation, model validation and prediction comes along with. Based on location

[1]https://www.channelnewsasia.com/news/asia/we-are-scared-deadly-dengue-outbreak-overwhelms-bangladesh-11792080

patterns, the data can be classified into two broad classes, such as regular or irregular grid. The regular grid point is also known as lattice. Analysis of lattice data has its paradigm and many developments came out in the last several decades.

We focus to handle the big non Gaussian data through generalized linear geostatistical modeling(GLGM) approach which is more general in nature. The lattice data can be fit too into the GLGM framework by some modification in the dependence structure. The major development in this domain done by Diggle *et al.* (1998), Diggle and Ribeiro (2007), and Diggle and Giorgi (2019). We expand the `farmer` approach for GLGM which necessarily covers binary and binomial logistic models, Poisson models. In the following section, we discuss the model formulation, estimation and inference procedure for GLGM under `farmer` approach platform.

4.2 Model formulation and estimation procedure

Let us recall the model (1.1) in our mind where the dependent variable $Y(s)$ were considered as continuous and to follow the Normal distribution. Now, instead we consider the non Gaussian case for $Y(s)$, where, $Y(s)$ is the count of an event could follow a binomial or Poisson model. Therefore, according to the GLGM formulation the observed response $Y(s)$ is mutually independent conditioned on $S(s)$ with conditional expectation, $\mu(s)$. The linear predictor is expressed as,

$$g(\mu(s)) = X(s)^\top \beta + S(s) + e(s), \tag{4.1}$$

where $g(\cdot)$ is known as the *link function*, $\mu(s)$ is the conditional expectaion of the response $Y(s)$, $X(s)$ is the vector of possible of covariates, β is vector of length p for the covariates, $S(s)$ is unobserved Gaussian processes through which the spatial dependence induced in the system. The specification of distribution of $S(s)$ is same as done in the previous chapters. $e(s)$ is zero-mean independent normally distributed random effects with common variance τ^2 which is also known as nugget effect. In the generalized linear model setting this term can be interpreted as the error in the response uncaptured by the predictor variables. Our target is to estimate the parameters associated with the model (4.1). Let us represent the parameter set into two as $\theta = (\varsigma, \tau^2)$ where $\varsigma = (\beta, \sigma^2, \phi)$, σ^2 and ϕ are the variance and scale parameters associated with spatial process $S(s)$. The joint distribution of $Y = (Y(s_1), \ldots, Y(s_n)), S = (S(s_1), \ldots, S(s_n))$ and $e = (e(s_1), \ldots, e(s_n))$, following the hierarchical structure can be presented as,

$$[Y, S, e; \varsigma, \tau^2] = [S; \varsigma] \times [e; \tau^2] \times [Y|S, e; \varsigma, \tau^2], \tag{4.2}$$

where the square bracket $[A]$ represents the probability distribution of A. Then the likelihood function for parameter $\theta = c(\varsigma, \tau^2)$ given the observed data $y = (y(s_1), y(s_2), \ldots, y(s_n))^\top$ is obtained by integrating out S from the (4.2), i.e.,

$$L(\theta = (\varsigma, \tau^2)) = \int [S; \varsigma] \times [e; \tau^2] \times [y|S, e; \varsigma, \tau^2] dS. \tag{4.3}$$

Unfortunately this (4.3) integrals does not have a closed form solution. This is always a big challenge to find likelihood in closed form for non Gaussian dependent data. However, there are alternatives. There are several alternative ways to estimate the parameters from the model (4.3).

A mention can be made of four approaches, (*i*) Hierarchical likelihood method (Lee and Nelder (1996)), (*ii*) Laplace approximation (Wolfinger (1993)), (*iii*) Monte Carlo sampling (Geyer and Thompson (1992)), and (*iv*) Generalized estimating equation (Zeger *et al.* (1988)). A summary of the processess can be found in Diggle and Ribeiro (2007) and Diggle and Giorgi (2019) and the references thereof. For convenience of reader we have echoed here the objective functions or equation to solve for all the four methods very briefly.

The hierarchical likelihood function is defined avoiding the integration with respect to S. The logarithm of hierarchical likelihood function is defined by Lee and Nelder (1996) as,

$$L_{HL}(\varsigma, \tau^2) = \log[S; \varsigma] + \log[e; \tau^2] + \sum_s \log[y(s)|S(s), e(s); \varsigma, \tau^2] \tag{4.4}$$

Laplace approximated log likelihood (4.3) based on second order Taylor expansion is obtained,

$$\log L_{la}(\varsigma, \tau^2) = \log[\hat{S}, y; \varsigma] - \frac{1}{2}|H(\hat{S})|, \tag{4.5}$$

where $\hat{S}$ is the maximized value of S and $H(S) = -\frac{\partial^2 \log[S,y;\varsigma,\tau^2]}{\partial^2 S}$, y is the observed data, ς and τ^2 are the vector of parameters including regression and spatial parameters, the subscript la means the Laplace approximation.

Monte Carlo sampling based approximation gives the likelihood(MCML),

$$L_{mc}(\varsigma, \tau^2) = \frac{1}{B}\sum_{k=1}^{B}\frac{[s^{(k)}, y; \varsigma, \tau^2]}{[s^{(k)}, y; \varsigma_0, \tau_0^2]}, \tag{4.6}$$

where B is the number of samples drawn, $s^{(k)}$ represents the realization of the vector S at k^{th} draw from the multivariate normal distribution with specified parameter values ς_0, τ_0^2, similarly the subscript mc represent the Monte Carlo.

There is the fourth method where instead of maximizing likelihood function Zeger *et al.* (1988) proposed to solve the estimating equation which is called *generalized estimating equation*(GEE). This primarily developed for analyzing the correlated longitudinal data. The β parameter can be estimated consistently solving the estimating equation,

$$\frac{\partial \mu}{\partial \beta}\Sigma^{-1}(Y - \mu) = 0 \tag{4.7}$$

where μ is the mean vector of Y which is a function of β, $\Sigma = \text{var}(Y)$ is function of mean μ. Through the off-diagonal element of Σ, the spatial dependence can be introduced. Later, Gotway and Stroup (1997) adopted this approach in spatial setting. The authors suggest estimating the regression parameters temporarily ignoring the spatial dependence and then estimate the spatial parameters by smoothing the variogram of standardized residuals.

Now, optimizing (4.4), (4.5) and (4.6) or adopting GEE approach we can estimate the parameters of our interests. However, all the methods do not perform similarly in every situation. Such as, the Laplace approximation performs well when the number of trials $n(s)$ at location s are large enough and the probability of success $p(s)$ at location s closer to boundary, on the other hand, the estimators obtained using MCML converges to the Maximum likelihood estimators when the number of draws increases to infinity (see Diggle and Giorgi (2019)). Based on researchers' interest the approach can be chosen. For time constraint we could not apply all four approaches in the current thesis.

In `farmer` framework we propose to implement any of the four approaches at block level to obtain the block estimates $\hat{\theta}_i$ and respective observed information matrix $J(\hat{\theta}_i)$. The unavailability of the Fisher information matrix is a limitation of the non Gaussian `farmer` approach. However, the observed information matrix is also a good approximation of Fisher's information and usefulness of it is argued by Efron and Hinkley (1978).

Due to the same reason, we will be unable to apply the bias correction technique at this point. This is a limitation however we have already brought this issue in our to-do list. The second stage modeling for combining local estimates and variance estimation has been proposed to do in the same fashion as done for the Gaussian case. Therefore, the description of the $MCAR$ model fitting and variance estimation procedure will not be repeated for the binomial case. As a replacement of the Fisher information matrix, the observed information matrix will be used.

The following section describes the `farmer` approach for logistic models.

4.3 Binomial data

This section is dedicated to model formulation, estimation and inference of GLGM for binomial count under the `farmer` framework.

Let us assume that the response $Y(s)$ comes from a binomial population with $n(s)$ be the number of trails and $p(s)$ is the probability of an event which is common to all the trails at location s. Then the link function in equation (4.1) is the *logit* link and the model is binomial logistic model for $Y(s)$ which can be expressed as,

$$logit(p(s)) = \log\left\{\frac{p(s)}{1-p(s)}\right\} = X(s)^\top\beta + S(s) + e(s), \tag{4.8}$$

where $e(s)$ has the zero-mean normal distribution with common variance τ^2, $X(s)$ is the possible vector of covariates and β is the associated vector coefficients, $S(s)$ is the latent process. Our target is to estimate the parameters associated with the model (4.8). This can be done replacing $[y(s)|S(s), e(s); \varsigma, \tau^2]$ by the Binomial probability mass function at location s. This will give us the likelihood function for binomial data. We can follow the algorithm (1) with slight modification as described in previous section for obtaining the model parameters and their variances.

In the next section, we present the output from a set of simulation experiments. We have designed the simulation experiments based on experience gathered from the Gaussian part. The detailed procedure is explained in the following subsection.

4.3.1 Simulated example for binomial data

In these simulation examples we have considered the simple model (4.8) with one single explanatory variable. We have assumed the Matérn covariance model for the spatial process $S(s)$. This covariance function has already been introduced in the first chapter

of this thesis. We have considered the true values are $\beta_0 = \beta_1 = \sigma^2 = 1,\ \phi = 0.15, \tau^2 = 0.1,$ and $\kappa = 0.5$, where κ is the smoothness parameter of the Matérn covariance model. The exponential covariance model is the special case of Matérn covariance model when $\kappa = 0.5$. The one single explanatory variable is considered which is known realization from $N(0, 0.5^2)$. We have generated two dimensional coordinates uniformly over the domain $(0, 30) \times (0, 30)$. To allow different block size we have done two different scenarios with $n = \{20,000,\ and\ 32,000\}$ locations which gives average block sizes 250 and 400 locations with 80 blocks. At each of these locations we have generated,

$$d_1(s) = 1 + X_1(s) + S(s) + e(s),$$

where $S(s)$ comes from Gaussian process with aforementioned parameter values. These generated realizations are then converted into probability scale using the following formula,

$$p(s) = \frac{\exp(d_1(s))}{1 + \exp(d_1(s))}$$

The $p(s)$ is now considered the true probability of success at s^{th} location for simulation. Considering this probability and the number of trials, $n(s) = 40$ we have generated a single binomial count at that location.

To estimate the model parameters and respective observed information we have applied the MCML method using the `PrevMap R` package available at the comprehensive R archive network (CRAN). Of-course other methods could also be applied however due to time constraints we could not do at this moment. This package allows us to estimate the parameters for Gaussian, binomial as well as Poisson model along with their observation information matrix with no difficulties. However, this taking a bit more time. The MCML, Laplace and low-rank approximation methods for binomial logistic and Poisson log-linear models are implemented in this package. There is also flexibility of practicing Bayesian inference and multivariate prediction through this package. The outputs of the experiments are presented in the box plots and table below. In the table, we have presented ten percent trimmed mean.

From the table (4.1) and figures (4.2, 4.3, 4.4, 4.5, 4.6) we see the similar behavior of `farmer` approach for estimating regression parameters as seen before for Gaussian regression case. They are well estimated for medium and large block sizes. All the three `farmer` estimators `bar, fem`, and `rem` perform equally which was observed for Gaussian spatial regression too. The scale parameter ϕ is underestimated by `fem` and `rem` however the `bar` captured the true value very well. For variance parameters, the

Scenarios	Pars(truth)	$\hat{\theta}_{\texttt{bar}}$	$\hat{\theta}_{\texttt{fem}}$	$\hat{\theta}_{\texttt{rem}}$
$K = 80,\ m \approx 250$	$\beta_0(1.00)$	1.0012	0.9996	0.9996
	$\beta_1(1.00)$	1.0004	0.9975	0.9975
	$\log(\sigma^2)(0.00)$	-0.0504	0.0092	0.0091
	$\log(\phi)(-1.897)$	-1.9076	-1.9709	-1.9708
	$\log(\tau^2)(-2.303)$	-7.1085	-1.7923	-1.7915
$K = 80,\ m \approx 400$	$\beta_0(1.00)$	1.0068	1.0055	1.0055
	$\beta_1(1.00)$	0.9996	0.9983	0.9983
	$\log(\sigma^2)(0.00)$	-0.0417	-0.0057	-0.0059
	$\log(\phi)(-1.897)$	-1.9141	-1.9742	-1.9740
	$\log(\tau^2)(-2.303)$	-3.2868	-1.7772	-1.7763

TABLE 4.1: Estimates obtained from the set of experiments for binomial data using `farmer` approach.

performance is the opposite. Both the σ^2 and τ^2 are estimated well closer to true values by `fem` and `rem` however the `bar` is underestimated. The nugget parameter is always harder to estimate and which is happened here too. The `bar` is underestimated and `fem, rem` is bit overestimated. One important observation is that for the spatial binomial model, the performance of all three estimators for any parameter does not improve after increasing the block size from 250 to 400. This could be due to several reasons. Firstly, the likelihood is not maximized using the exact method rather has been approximated using MCML. The MCML estimates tend to MLE when the number of draws gets larger. Therefore, if MLE's are underestimated or overestimated then it less probable to cover the true value by this approach. The approximation methods have asymptotic properties however at block level that may not always be true. Also, the behavior of approximation methods depends on the low and high value of parameters of the binomial distribution. Other approximation methods could be tested to conclude. Secondly, due to unavailability of the closed form of likelihood, we could not estimate the Fisher's information matrix. Instead, we have used the observed information matrix.

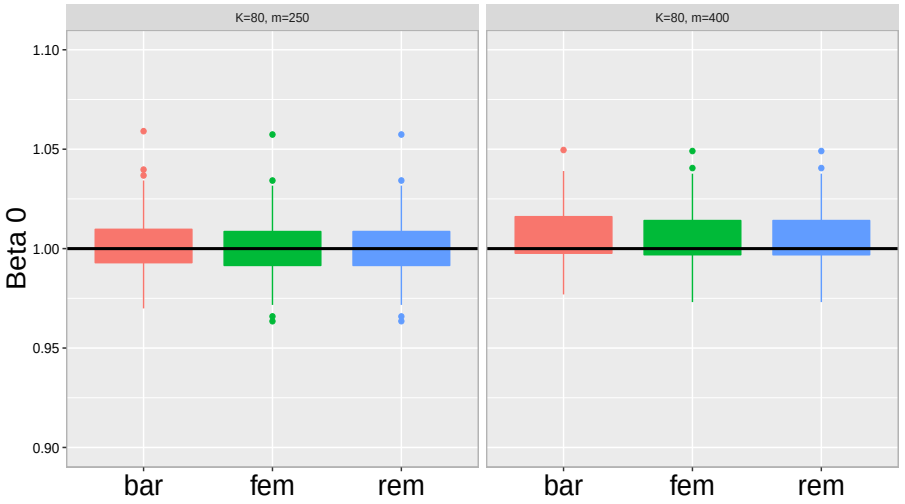

FIGURE 4.2: Box-plots of the farmer estimates of β_0 for $(K \times m) = (80 \times 250)$, *and* (80×400) in column 1 and 2.

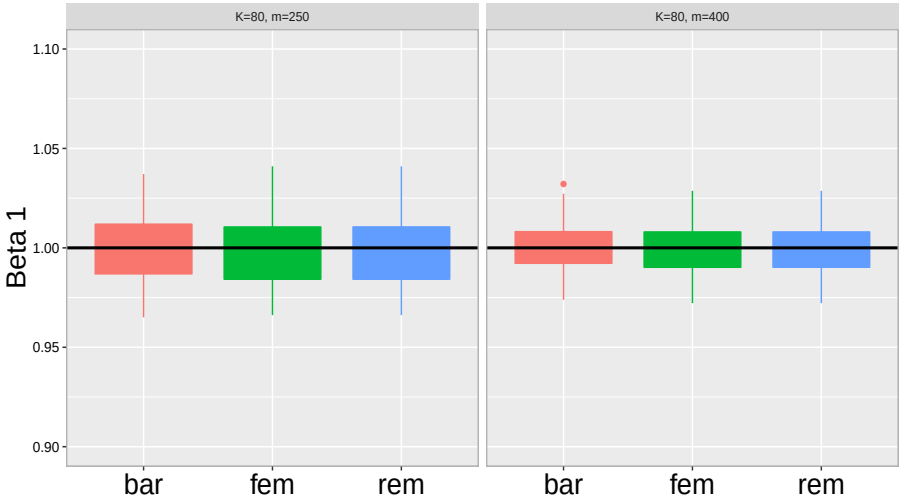

FIGURE 4.3: Box-plots of the farmer estimates of β_1 for $(K \times m) = (80 \times 250)$, *and* (80×400) in column 1 and 2.

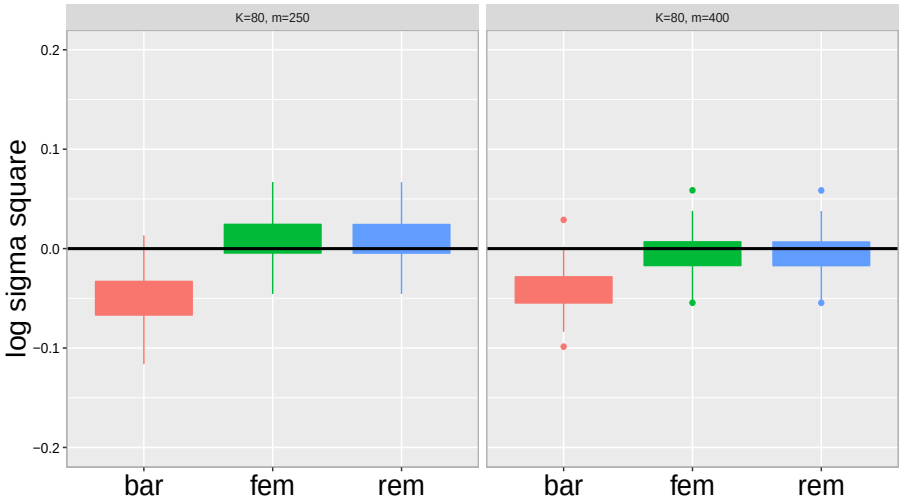

FIGURE 4.4: Box-plots of the farmer estimates of $\log(\sigma^2)$ for $(K \times m) = (80 \times 250)$, *and* (80×400) in column 1 and 2.

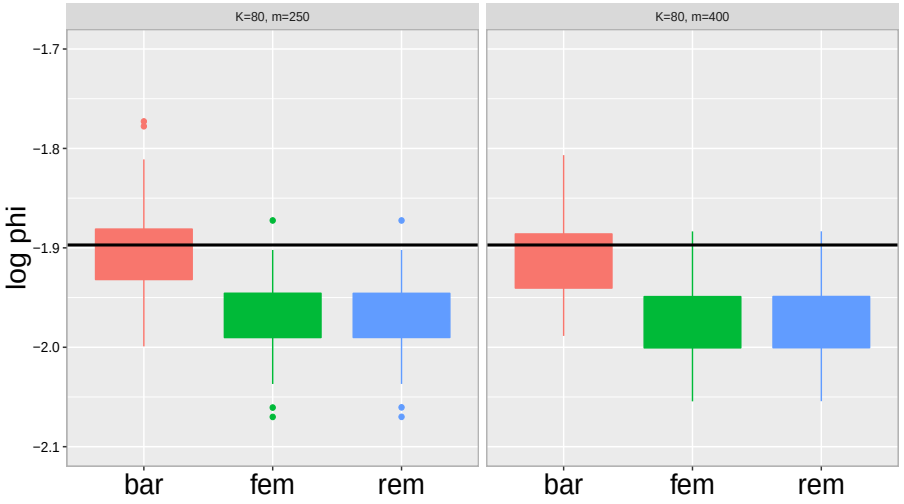

FIGURE 4.5: Box-plots of the farmer estimates of $\log(\phi)$ for $(K \times m) = (80 \times 250)$, *and* (80×400) in column 1 and 2.

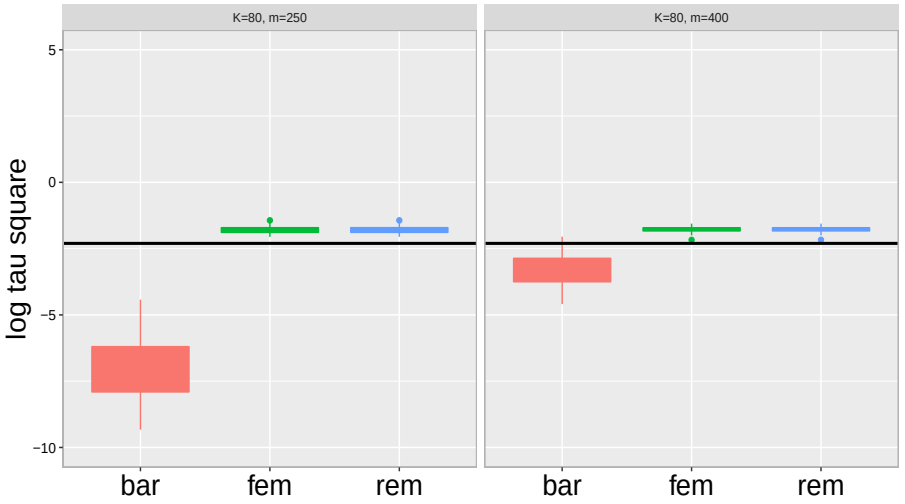

FIGURE 4.6: Box-plots of the farmer estimates of $\log(\tau^2)$ for $(K \times m) = (80 \times 250)$, *and* (80×400) in column 1 and 2.

The contrasting behavior of `farmer` estimators in estimating variance parameter σ^2 and scale parameter ϕ could be explained by the following figure (4.7). In the following plot, we have plotted the $\log \sigma^2$ versus $\log \phi$. The association is negative for every case though the strength of the association varies. This is suggestive that if $\log \sigma^2$ is overestimated then there is a possibility that the $\log \phi$ will be underestimated or vice versa.

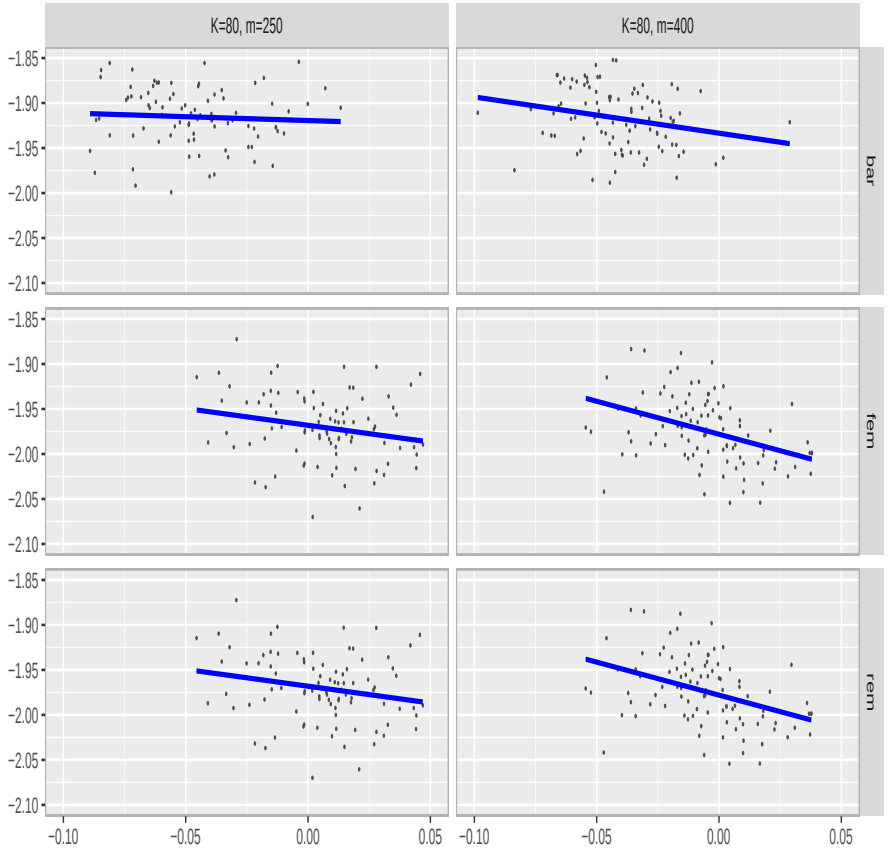

FIGURE 4.7: Scatter plots of $\log(\phi)$ (y-axis) and $\log(\sigma^2)$ (x-axis) for different scenarios and type of estimators.

In the following figures (4.8, 4.9, 4.10, 4.11 and 4.12) we have presented the confidence interval (CI) of estimated parameters of the model (4.8). The `farmer bar, fem` and `rem` produce very small standard error for the regression parameters. The length of 95% CI's are similar among the three versions of estimators however the `rem` expected to provide a bit wider CI than others because this estimator accounts for the between block dependence. The CI for `rem` is very slightly wider for smaller block size however this is hardly visible with open eyes for two regression parameters β_0 and β_1. On the other hand for spatial dependence parameters $\sigma^2, \phi,$ and τ^2 the difference is visible. `fem` and `rem` produce wider CI than that of `bar`. Also, the estimates of these three parameters produce comparatively larger standard errors where the estimate of the nugget parameter produces the largest one. For both the σ^2 and τ^2 the variance reduces as the sample size increases while for ϕ that is not the case. This is an indication that the `farmer` estimators have empirical infill asymptotic behavior for some parameters. However, this is empirical evidence only, the mathematical investigation is required to conclude. If we could use the Fisher information matrix instead of the observed

information matrix in the `farmer` algorithm the bias correction techniques could be applied. This would improve performance.

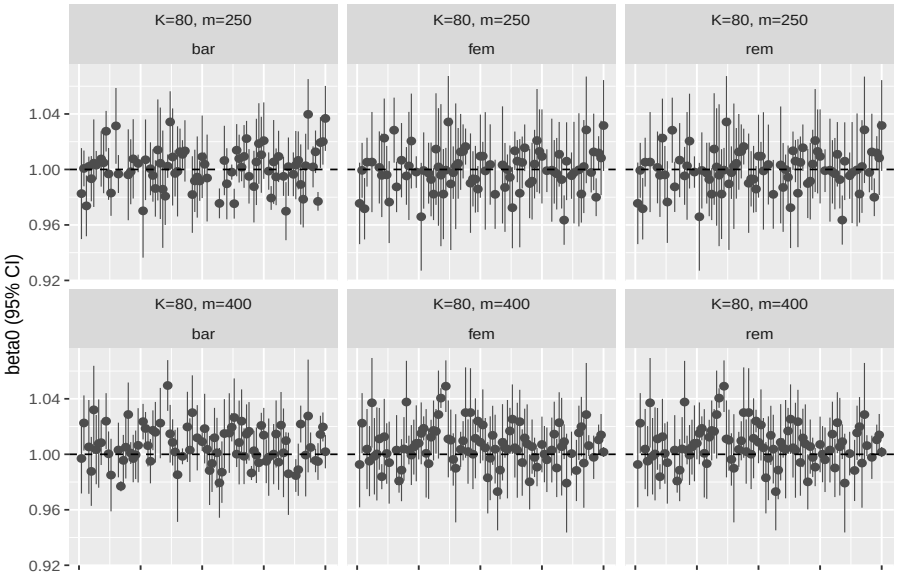

FIGURE 4.8: Estimate for β_0 and their confidence interval.

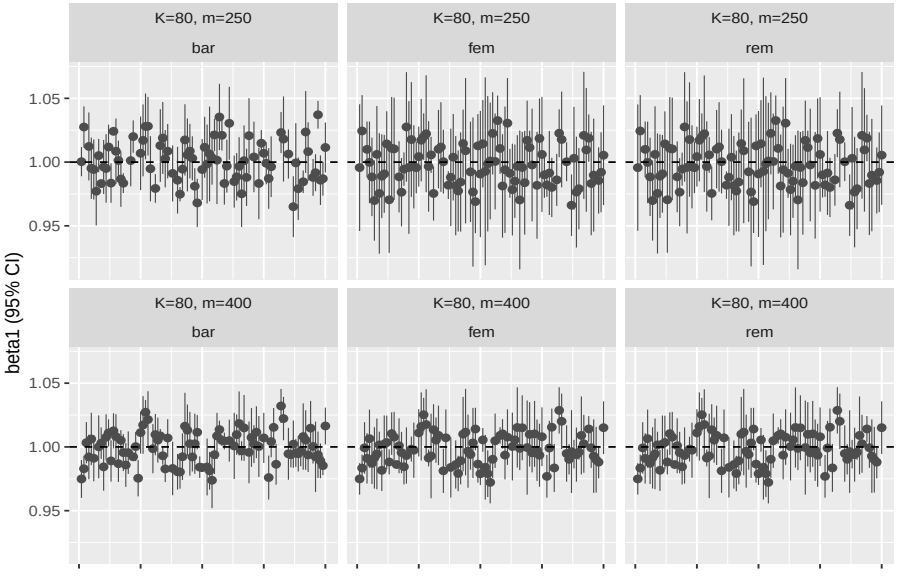

FIGURE 4.9: Estimate for β_1 and their confidence interval.

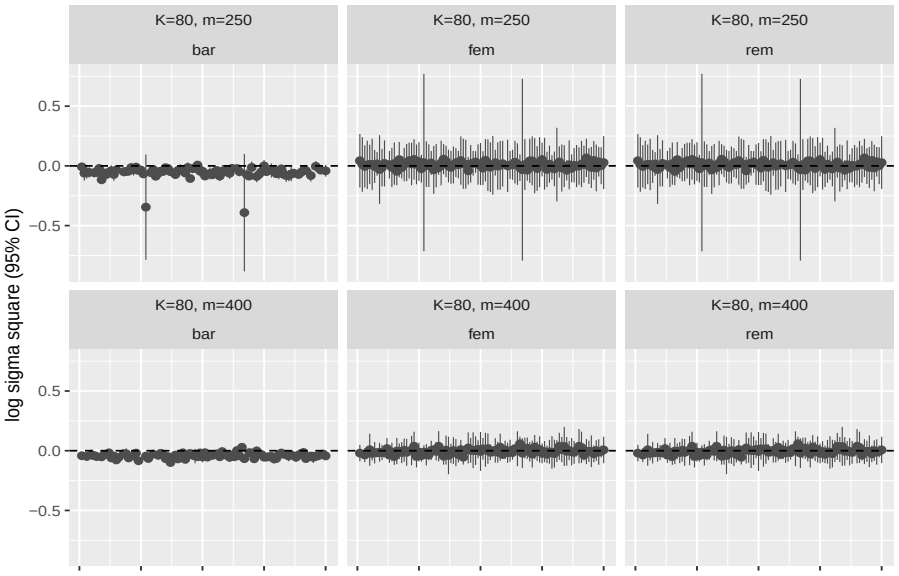

FIGURE 4.10: Estimate for $\log \sigma^2$ and their confidence interval.

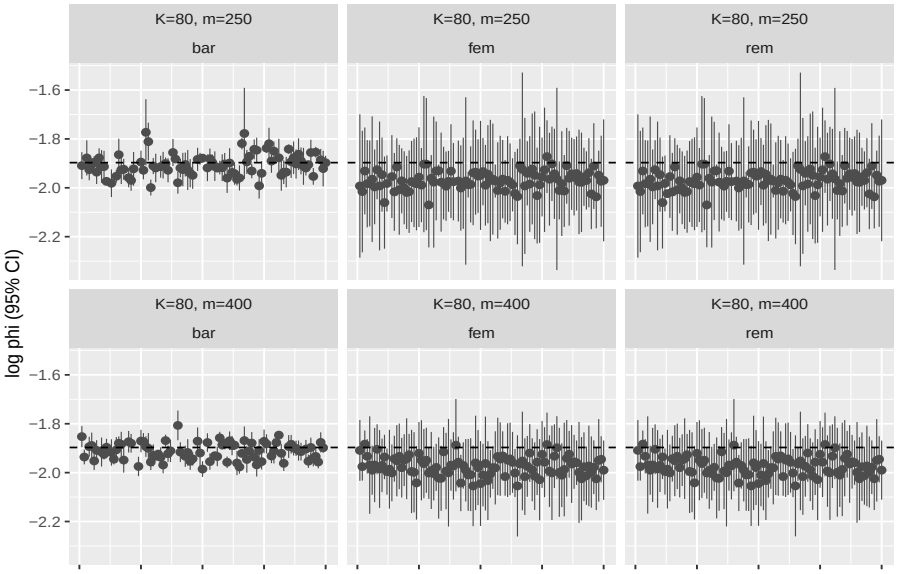

FIGURE 4.11: Estimate for $\log \phi$ and their confidence interval.

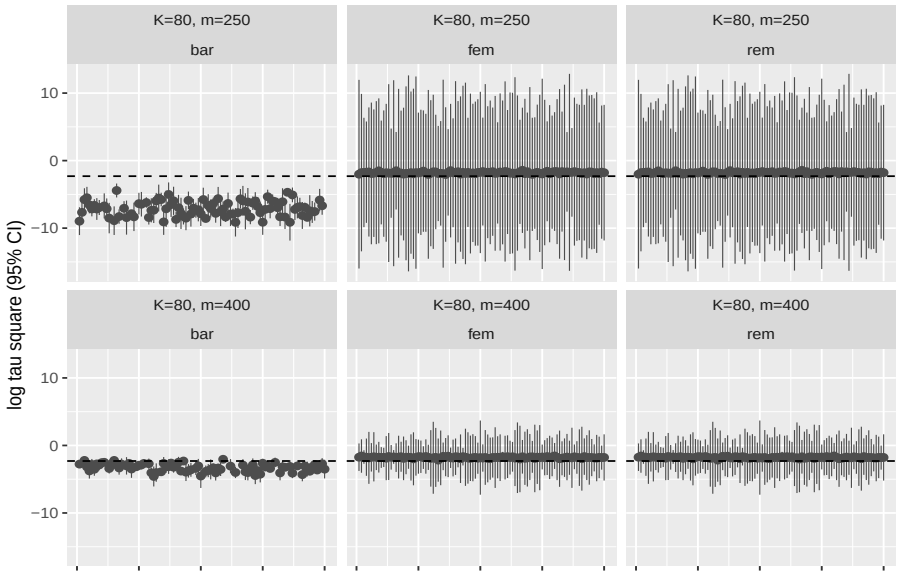

FIGURE 4.12: Estimate for $\log \tau^2$ and their confidence interval.

Another important observation is that the CI of `bar, fem` and `rem` for σ^2 and τ^2 become closer in length as the number of observation increase inside the block. This is suggestive that if the block size is sufficiently large then blocks generate similar information matrix and that is why the weight becomes almost unit for every block. In the following subsection, we present the application of the spatial binomial regression model to the same African data presented in section (3.4).

4.3.2 Application to African river blindness data

The detail description of the data set can be found in section (3.4). Since in the data we do not have any covariate we have considered the simple binomial logistic model (4.8) excluding the covariate. The same blocking strategy is used as done in section (3.4). The difference is that in the previous section we have used the Gaussian model after logit transformation of the count but here we have used GLGM with *logit* link for family binomial. In the previous section, we have not presented the blocks and their connections. In the figure (4.13) we have presented the constructed blocks and the between block connections. The connection is defined based on the distance 1000 kilometers(km). That is if centers of two blocks are within 1000 km they are assumed to be associated. The threshold is determined by trial and error with the objective that every adjacent block should be connected. This produces an irregular network. To accommodate the irregularity we have constructed the adjacency matrix W in a way

that the strength of association between two blocks varies based on the number of blocks they are associated. The row total of the matrix W is equal to 1.

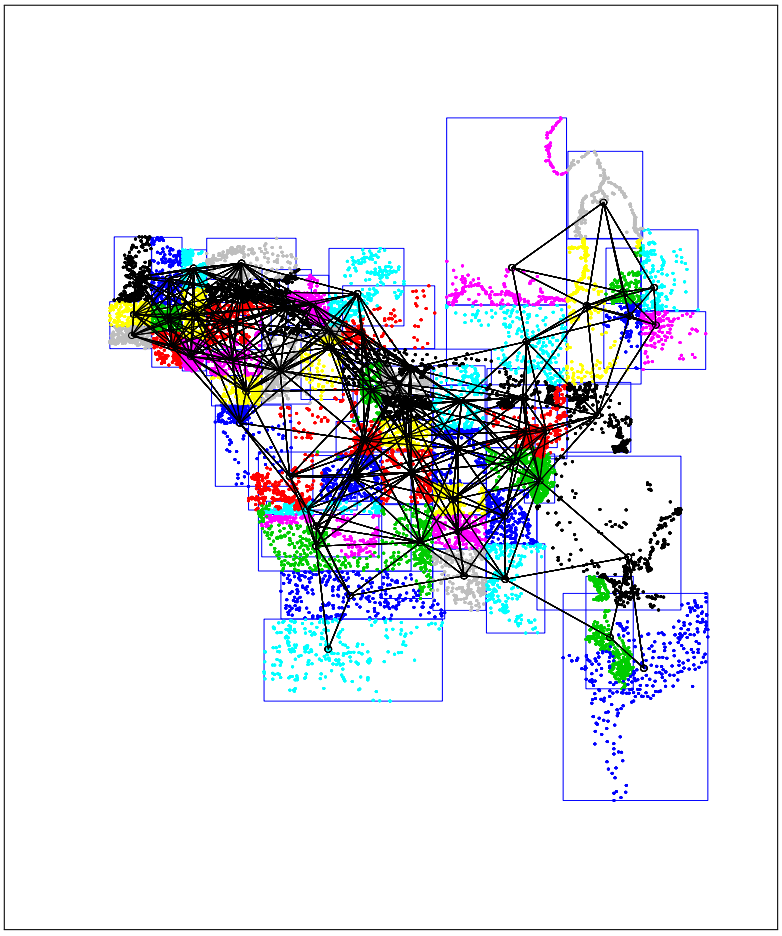

FIGURE 4.13: African river-blindness data blocks and network.

We have considered the Matérn covariance model for the spatial process with smoothness parameter $\kappa = 0.5$ and $\kappa = 2$. Where the first case gives us the parameter estimation of the exponential covariance model and the second case provides estimates for Matérn model. The second choice is following Noma *et al.* (2014) for comparison purposes. They applied the Laplace approximation for estimating the parameters of the model with Matérn covariance model with $\kappa = 2$. We have employed the Laplace approximation method for estimating the parameters and observed information matrix at block level using `PrevMap` package. The results are presented in figure (4.14).

In the figure we have plotted the three point estimates `bar, fem` and `rem` along with their 95% CI's obtained for exponential and Matérn model. We have also plotted the MLE obtained using the entire data set as reference. The MLE is taken from Noma *et al.* (2014).

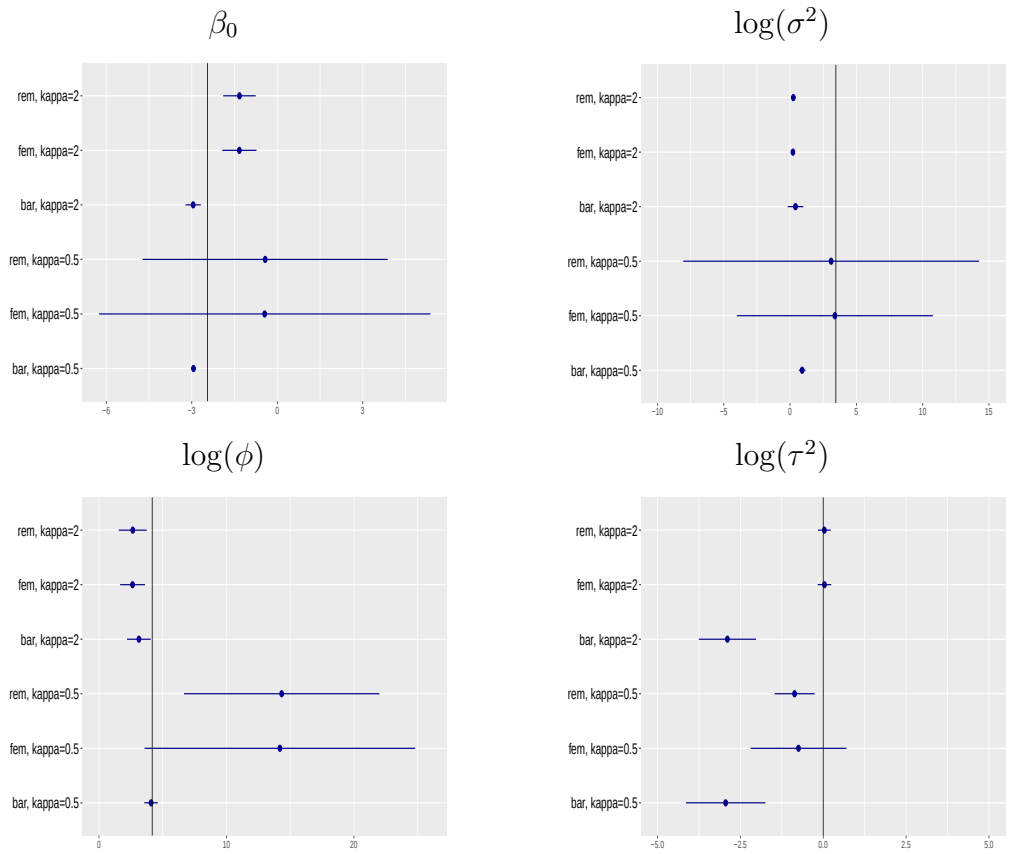

FIGURE 4.14: African river-blindness data: comparison of various estimates with MLE(vertical solid black line).

It seems hard to reach a discrete conclusion about the results. Estimates are showing varying behavior by type. However, estimates for mean β_0, variance σ^2 and scale ϕ show similar behavior. For, $\kappa = 2$ they are quite closer to the MLE. In case of nugget effect `fem` and `rem` are very close to the reference line while `bar` is not. The behavior of `farmer` estimators in simulation experience are reflected here in real the example too. The CI in case of Matérn model is too narrow for all versions of estimators however that is very wide for `fem` and `rem` in case of the exponential model. Based on these results we may recommend this approach however extra care should be taken in selecting an appropriate covariance model. This should be a challenge still now. Unfortunately, at this point, we are unable to suggest any approach to select the covariance model. However, a possibility could be to estimate the smoothness parameter of Matérn family in `farmer` fashion through variogram estimation.

This is not exactly the "end" however have to stop now. Further, developments needed in this area aiming to find ways to incorporate the bias reduction method, improving the information matrix estimation, data-based selection of likelihood approximation approach and selecting the appropriate covariance model. The generalized

estimating equation approach could be a possible way to incorporate the bias reduction technique. In the next section, we mention some possible applications of `farmer` approach where big spatial data is a big challenge to handle.

4.4 Other potential applications of `farmer` approach

The `farmer` approach is a general framework and has the potentials to apply in diverse fields including both spatial and nonspatial data. Up to this point we have shown application of `farmer` approach in *geostatistics*. However, the application of the approach is not limited to that specific field. In this section, we show some possible application of the proposed approach in other scenarios.

4.4.1 Poisson log linear model

Another widely used non Gaussian data is open ended count. Let us assume the $Y(s)$ is the count of the target event at location s with expected value $\lambda(s)$. Then using log link the Poisson *log linear* model is expressed as,

$$\log \lambda(s) = X(s)^\top \beta + S(s) + e(s), \tag{4.9}$$

where $\beta, X(s), S(s),$ and $e(s)$ have similar interpretation as in previous section. To estimate the parameters of the model we propose to follow the similar procedure as proposed for binomial data. However, in case of Poisson log linear model we have to replace $[y(s)|S(s), e(s); \varsigma, \tau^2]$ with the Poisson distribution for the s^{th} location. We can follow the algorithm (1) with modification described for binomial case for obtaining the model parameters and their variances.

One characteristic of Poisson distribution is that the mean is equal to its variance. However, this is not guaranteed always rather there are many cases where the data generated variance is larger than the mean. This is called extra-Poisson variance or *overdispersion*. The last term in the model (4.9) is some times interpreted as the combination of both extra-Poisson variance and the effect of unmeasured covariates however they are not easily distinguishable.

4.4.2 *Ising* model estimation

Besag (1974) discussed spatial interaction and the statistical analysis of lattice systems in his seminal papers utilizing the Hammersley-Clifford theorem and Markovian properties. He suggested the *auto-models*, such as *auto-normal*, *auto-logistic* and *auto-binomial* for off and on lattice systems. Based on *auto-models* the full conditional distribution for binary and count data can be constructed feasibly. The *auto-logistic* or binary Markov Random Field (MRF) is basically same as the Ising model which primarily developed for studying the *ferromagnetism* in physics (see Ising (1925)). The Besag (1974)'s work has opened the window for analyzing the non-Gaussian spatial data however the methods are computationally expensive. The estimation of the *auto-logistic* model involves summation over 2^n terms, where n is the number of data points. Even for a low medium size sample, such as, on a 10×10 lattice its a challenge. Several authors have proposed solutions based on approximating likelihood,, recursive algorithm, approximating pseudo likelihood to overcome this challenge (Bartolucci and Besag (2002), Hardouin and Guyon (2014), Tjelmeland and Austad (2012)). These approaches can handle medium size sample. Therefore, big binary spatial data handling is still an open challenge. `farmer` approach could be a hope to overcome this challenge.

Let us assume $Y = (Y_1, Y_2, \ldots, Y_n)$ is a binary random vector associated with $n-$ locations. A snap of realization of the Y is presented in figure (4.15). In this figure, we notice that the 6^{th} cell (red in color) is connected with cells (blue in color). A *clique* is a set of locations in which every location is a neighbor of all other locations in the set or contains a single element only. For simplicity we will assume the *clique* of size $k = 2$ which gives us first order dependence. Adding second or more order dependence have not much benefits rather invites complications. Let us denote the conditional probability of success at location s is $\pi(s)(\cdot)$ which is defined as,

$$\pi(s)(\cdot) = P(Y(s) = 1|\ y(s^{'}); s^{'} \neq s) \tag{4.10}$$

For a clique of size $k = 2$ the $y(s^{'}) = (y(s+1, s^{'}), y(s-1, s^{'}), y(s, s^{'}+1), y(s, s^{'}-1))$. From figure (4.15), the red cell will depend on only the blue cells.

1	0	0	1
0	1	1	1
1	1	0	0
0	1	1	1

FIGURE 4.15: Example of binary spatial data on regular lattice

The probability distribution in equation (4.10) is constructed by Besag (1974) and later Carlin *et al.* (2014) described in details with recent developments and other possibilities. The full conditional is defined as

$$\pi(s)(\cdot) = \frac{e^{\psi S(s,1)}}{e^{\psi S(s,1)} + e^{\psi S(s,0)}}, \tag{4.11}$$

where $S(s,1) = \sum_{s \sim s'} 1(y(s') = 1)$ and $S(s,0) = \sum_{s \sim s'} 1(y(s') = 0)$ and ψ controls the weight on matching as described in the aforementioned book. We can easily obtain the logit function for $\pi(s)(\cdot)$ as,

$$\log\left\{\frac{\pi(s)(\cdot)}{1-\pi(s)(\cdot)}\right\} = logit(\pi(s)) = \psi(S(s,1) - S(s,0)). \tag{4.12}$$

and for regression setting with vector of covariates $X(s)$ the model in equation (4.12) becomes,

$$logit(\pi(s)) = \psi(S(s,1) - S(s,0)) + X^{\top}(s)\beta \tag{4.13}$$

where, β is the vector of regression coefficients. Now, we have two unknown parameters (ψ, β) needs to be estimated.

This model is computationally intractable for medium size data. We can apply the `farmer` algorithm to handle the large binary lattice data.

This is not the least there could be other spatial and nonspatial area where the

`farmer` approach can be applied suitably. Such as point pattern analysis, image analysis, functional data analysis, etc. The main task is to define a suitable model at the block level and estimate them appropriately. The quality of estimators in the proposed approach greatly depends on the quality of local estimates. In the next chapter, we point out some concluding remarks and limitations. Also, there is a lot to do therefore we discuss briefly the way forward.

Chapter 5

Conclusions and way forward

All scientific work is incomplete
—Austin Bradford Hill (1897-1991)

5.1 Concluding remarks

All the mainstream disciplines have experienced data explosion due to technological advancements in recent times. Many of these data are location associated and they are called *spatial data*. The volume, veracity, variety, and velocity along with the dependence structure of spatial data have made the analysis more challenging. To overcome this challenge we have developed a new divide and conquer approach which we call `farmer` approach. In this method, we propose to split the data into mutually exclusive blocks and estimate the block summaries by some method such as the maximum likelihood estimation procedure. Block summaries are combined using fixed and random effect meta analysis models. The estimators that we obtain are called `fem` and `rem` estimators. This is a general platform by which many big data problems can be dealt with.

The proposed approach is intuitive, easy to implement and computationally efficient. This is multiple times faster than the novel maximum likelihood approach and a good competitor in the same domain. The `farmer` approach is suitable to implement in parallel. Parallelization allows conducting very very big computation faster using multi-threading. Therefore, the approach has potentials in the big data computations industry who use multiple machines. The performance of the estimates in the subsampling approach depends on the blocking strategy. This is a common pitfall of this type of approach. Our method is not free from this objection. Finding an optimal blocking strategy could be cumbersome.

The `farmer` approach provides a more realistic measure of the standard error of the

estimate. Not only reducing the computational burden but also estimating standard error is equally important. This contribution will enable assessing the quality of estimates in large data modeling. When multiple machines are employed where some are nearer and some are distant this approach can help to assess the quality of estimates.

Downward bias in the subsampling approach is a known problem. Adjusting score function using Jeffreys prior penalty function at block level reduces the bias by a great amount. This is an advancement over existing approaches. Though the bias correction method was proposed for independent data however working nicely for dependent data as well. The bias correction method applied here in the `farmer` approach does not have the invariance properties. That is for the re-parameterized model there could have some disturbance in standard error estimate.

Consistency is a question in the estimation of the spatial model, especially for the spatial dependence parameters. Our approach provides estimates that are empirically consistent under infill as well as increasing-domain asymptotics. However, to confirm this a proper mathematical investigation is required. The random quantity included in the farmer estimators could be a hurdle to prove the asymptotic properties.

Another important advancement from our method is proposing a way to deal with large non-Gaussian data. Non-Gaussian spatial data is computationally intractable. Accommodating approximation techniques the `farmer` approach can handle large data faster with some limitations. The bias correction technique was not possible to apply therefore the dependence parameters found to be biased in simulation experiments. We need to find a way to apply the bias correction method for nonnormal data.

The theory behind the approach is excellent, novel techniques are utilized however evaluation of the method has not been sufficient yet. More rigorous simulation experiments are required with regular and irregular points settings. We have tested only simple models with a single covariate only while in real life this is not always the case. The prediction is a major interest in spatial data analysis which we have not dealt with yet. Another lacking is that no guideline is provided to deal with the non-stationary process. Not only our method but many other methods in the domain of big spatial data computation suffer from this drawback. In the next section we present some future research directions surrounding `farmer` approach.

5.2 Way forwards

Firstly, we would like to fill the gaps by mathematically investigating the asymptotic properties of the `farmer` estimators, adopting bias correction technique for non normal

data and conducting more rigorous simulation experiments. The presence of randomness in weight involved in estimators could make the mathematical investigation harder. Need to find a way out. Application of generalized estimating equation approach at block level for nonnormal data could be a possible way to adopt the bias correction technique. This could be done estimating the mean parameters ignoring the spatial dependence and then estimate the spatial dependence parameters from the standardized residuals. In the second stage, residuals can be assumed to come from the Gaussian process. This will allow us to adopt the bias correction method for spatial dependence parameters. Finding out optimal blocking strategy is also a target.

Secondly, the proposed approach possibly be extendable for spatial point pattern analysis. Spatial point pattern is useful for many scientific questions in forestry, oceanography and chemical research. Another challenge in the spatial arena is Besag (1974)'s auto models estimation for large data. Splitting the domain into smaller subdomains `farmer` approach would be an optimal solution for this.

Appendix

We consider the model (3.1) for simulation experiments. We have considered the number of blocks to $K = 80$, average block size $m \approx 250$, the total number of locations is $n = 20000$ over the domain $(0, 30) \times (0, 30)$. We have included a single explanatory variable in the linear spatial regression model which is known and generated from $N(0, 0.5^2)$. The true parameter values are considered as $\beta_0 = \beta_1 = \sigma^2 = 1,\ \phi = 2.0,\ \tau^2 = 0.1$. The results are presented in the figures below.

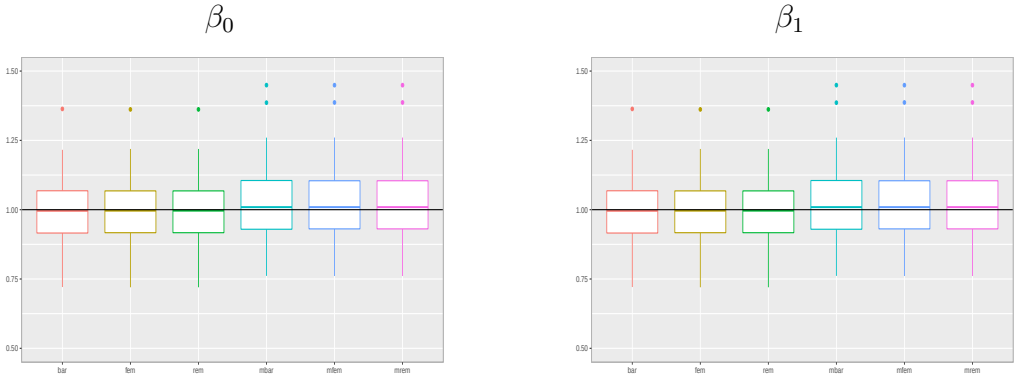

FIGURE .1: Boxplots of the bar, fem, rem, for β_0 and β_1. In each panel, the first three boxes are bias-uncorrected and the last three are mean bias-corrected ones.

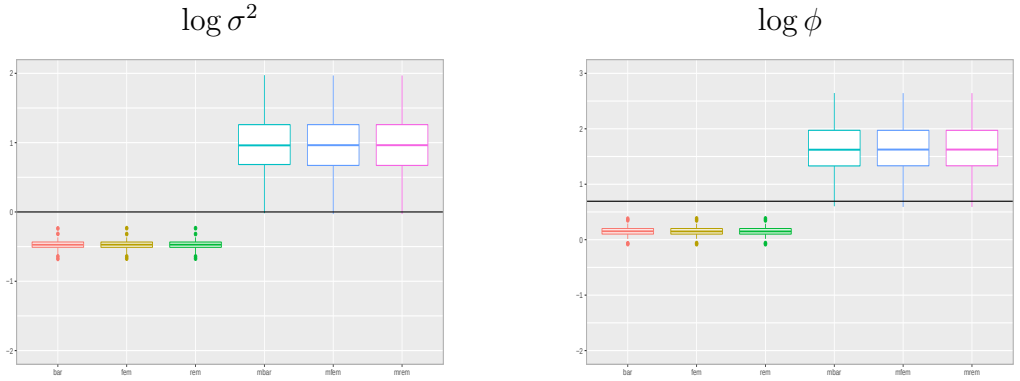

FIGURE .2: Boxplots of the bar, fem, rem, for $\log \sigma^2$ and $\log \phi$. In each panel, the first three boxes are bias-uncorrected and the last three are mean bias-corrected ones.

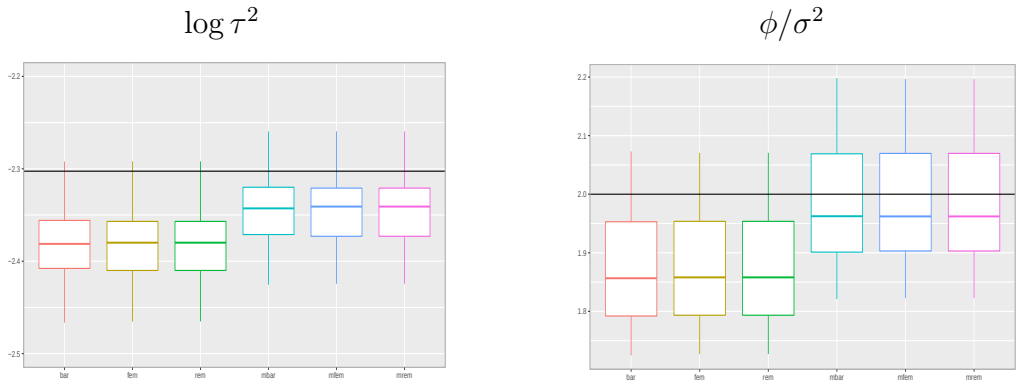

FIGURE .3: Boxplots of the bar, fem, rem, for $\log \tau^2$ and the ratio of ϕ/σ^2. In each panel, the first three boxes are bias-uncorrected and the last three are mean bias-corrected ones.

Bibliography

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2014) *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008) Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(4), 825–848.

Barbian, M. H. and Assunção, R. M. (2017) Spatial subsemble estimator for large geostatistical data. *Spatial Statistics* **22**, 68–88.

Bartolucci, F. and Besag, J. (2002) A recursive algorithm for markov random fields. *Biometrika* **89**(3), 724–730.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 192–236.

Bevilacqua, M., Gaetan, C., Mateu, J. and Porcu, E. (2012) Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach. *Journal of the American Statistical Association* **107**(497), 268–280.

Bickel, P. J., Götze, F. and van Zwet, W. R. (2012) Resampling fewer than n observations: gains, losses, and remedies for losses. In *Selected Works of Willem van Zwet*, pp. 267–297. Springer.

Carlin, B. P., Gelfand, A. E. and Banerjee, S. (2014) *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Chang, X., Lin, S.-B., Wang, Y. *et al.* (2017) Divide and conquer local average regression. *Electronic Journal of Statistics* **11**(1), 1326–1350.

Cressie, N. (1992) Statistics for spatial data. *Terra Nova* **4**(5), 613–617.

Cressie, N. and Johannesson, G. (2008) Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(1), 209–226.

Curriero, F. C. and Lele, S. (1999) A composite likelihood approach to semivariogram estimation. *Journal of Agricultural, biological, and Environmental statistics* pp. 9–28.

Dean, J. and Ghemawat, S. (2008) Mapreduce: simplified data processing on large clusters. *Communications of the ACM* **51**(1), 107–113.

Diggle, P. J. and Giorgi, E. (2019) *Model-based Geostatistics for Global Public Health: Methods and Applications*. Chapman and Hall/CRC.

Diggle, P. J. and Ribeiro, P. J. (2007) Model-based geostatistics. Springer.

Diggle, P. J., Tawn, J. A. and Moyeed, R. (1998) Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **47**(3), 299–350.

Efron, B. and Hinkley, D. V. (1978) Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika* **65**(3), 457–483.

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M. and Niemi, J. (2014) Estimation and prediction in spatial models with block composite likelihoods. *Journal of Computational and Graphical Statistics* **23**(2), 295–315.

Firth, D. (1993) Bias reduction of maximum likelihood estimates. *Biometrika* **80**(1), 27–38.

Fuentes, M. (2007) Approximate likelihood for large irregularly spaced spatial data. *Journal of the American Statistical Association* **102**(477), 321–331.

Furrer, R., Genton, M. G. and Nychka, D. (2006) Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics* **15**(3), 502–523.

Gelfand, A. E. and Vounatsou, P. (2003) Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* **4**(1), 11–15.

Geyer, C. J. and Thompson, E. A. (1992) Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society: Series B (Methodological)* **54**(3), 657–683.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. and Brilliant, L. (2009) Detecting influenza epidemics using search engine query data. *Nature* **457**(7232), 1012.

Gotway, C. A. and Stroup, W. W. (1997) A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural, Biological, and Environmental Statistics* pp. 157–178.

Grenander, U. and Rosenblatt, M. (1957) Statistical analysis of stationary time series .

Grenander, U. and Rosenblatt, M. (2008) *Statistical analysis of stationary time series*. Volume 320. American Mathematical Soc.

Grover, V., Chiang, R. H., Liang, T.-P. and Zhang, D. (2018) Creating strategic business value from big data analytics: A research framework. *Journal of Management Information Systems* **35**(2), 388–423.

Guhaniyogi, R. and Banerjee, S. (2018) Meta-kriging: Scalable bayesian modeling and inference for massive spatial datasets. *Technometrics* **60**(4), 430–444.

Hall, P., Horowitz, J. L. and Jing, B.-Y. (1995) On blocking rules for the bootstrap with dependent data. *Biometrika* **82**(3), 561–574.

Hardouin, C. and Guyon, X. (2014) Recursions on the marginals and exact computation of the normalizing constant for gibbs processes. *Computational Statistics* **29**(6), 1637–1650.

He, S. (2018) From beautiful maps to actionable insights: Introducing kepler.gl, ubers open source geospatial toolbox.

Hedges, L. V. and Vevea, J. L. (1998) Fixed-and random-effects models in meta-analysis. *Psychological methods* **3**(4), 486.

Hirano, T. and Yajima, Y. (2013) Covariance tapering for prediction of large spatial data sets in transformed random fields. *Annals of the Institute of Statistical Mathematics* **65**(5), 913–939.

Ising, E. (1925) Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik* **31**(1), 253–258.

Jin, X., Carlin, B. P. and Banerjee, S. (2005) Generalized hierarchical multivariate car models for areal data. *Biometrics* **61**(4), 950–961.

Johns, C. J., Nychka, D., Kittel, T. G. F. and Daly, C. (2003) Infilling sparse records of spatial fields. *Journal of the American Statistical Association* **98**(464), 796–806.

Jordan, M. I. *et al.* (2013) On statistics, computation and scalability. *Bernoulli* **19**(4), 1378–1390.

Kaufman, C. G., Schervish, M. J. and Nychka, D. W. (2008) Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association* **103**(484), 1545–1555.

Kelejian, H. H. and Prucha, I. R. (2007) Hac estimation in a spatial framework. *Journal of Econometrics* **140**(1), 131–154.

Kettenring, J. R. (1997) Shaping statistics for success in the 21st century. *Journal of the American Statistical Association* **92**(440), 1229–1234.

Kosmidis, I., Guolo, A. and Varin, C. (2017) Improving the accuracy of likelihood-based inference in meta-analysis and meta-regression. *Biometrika* **104**(2), 489–496.

Kosmidis, I., Pagui, E. C. K. and Sartori, N. (2018) Mean and median bias reduction in generalized linear models. *arXiv preprint arXiv:1804.04085* .

Kyriakou, S., Kosmidis, I. and Sartori, N. (2018) Median bias reduction in random-effects meta-analysis and meta-regression. *Statistical methods in medical research* p. 0962280218771717.

Lahiri, S. N. (2013) *Resampling methods for dependent data*. Springer Science & Business Media.

Lee, Y. and Nelder, J. A. (1996) Hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(4), 619–656.

Liang, F., Cheng, Y., Song, Q., Park, J. and Yang, P. (2013) A resampling-based stochastic approximation method for analysis of large geostatistical data. *Journal of the American Statistical Association* **108**(501), 325–339.

Liu, D., Liu, R. Y. and Xie, M. (2015) Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. *Journal of the American Statistical Association* **110**(509), 326–340.

Liu, Q., Bhadra, A. and Cleveland, W. S. (2018) Divide and recombine for large and complex data: Model likelihood functions using mcmc. *arXiv preprint arXiv:1801.05007* .

Mayer-Schönberger, V. and Cukier, K. (2013) Big data: A revolution that transforms how we work, live, and think.

Noma, M., Zouré, H. G., Tekle, A. H., Enyong, P. A., Nwoke, B. E. and Remme, J. H. (2014) The geographic distribution of onchocerciasis in the 20 participating countries of the african programme for onchocerciasis control:(1) priority areas for ivermectin treatment. *Parasites & vectors* **7**(1), 325.

Priestley, M. (1964) The analysis of two-dimensional stationary processes with discontinuous spectra. *Biometrika* **51**(1/2), 195–217.

Rue, H. and Held, L. (2005) *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC.

RUE, H. and Tjelmeland, H. (2002) Fitting gaussian markov random fields to gaussian fields. *Scandinavian journal of Statistics* **29**(1), 31–49.

Smith, T. E. (2009) Estimation bias in spatial models with strongly connected weight matrices. *Geographical Analysis* **41**(3), 307–332.

Stein, M. L. (1999) Interpolation of spatial data: some theory for kriging .

Stein, M. L. (2013) Statistical properties of covariance tapers. *Journal of Computational and Graphical Statistics* **22**(4), 866–885.

Tjelmeland, H. and Austad, H. M. (2012) Exact and approximate recursive calculations for binary markov random fields defined on graphs. *Journal of Computational and Graphical Statistics* **21**(3), 758–780.

Viechtbauer, W. (2005) Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics* **30**(3), 261–293.

Wikle, C. K. and Cressie, N. (1999) A dimension-reduced approach to space-time kalman filtering. *Biometrika* **86**(4), 815–829.

Wolfinger, R. (1993) Laplace's approximation for nonlinear mixed models. *Biometrika* **80**(4), 791–795.

Yang, H.-c., Dasdan, A., Hsiao, R.-L. and Parker, D. S. (2007) Map-reduce-merge: simplified relational data processing on large clusters. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pp. 1029–1040.

Yu, B. (2014) Let us own data science. *IMS Bulletin Online* **43**(7).

Zeger, S. L., Liang, K.-Y. and Albert, P. S. (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics* pp. 1049–1060.

Zhang, Y., Duchi, J. and Wainwright, M. (2015) Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research* **16**(1), 3299–3340.

Zhou, L. and Song, P. X.-K. (2017) Scalable and efficient statistical inference with estimating functions in the mapreduce paradigm for big data. *arXiv preprint arXiv:1709.04389* .

Zouré, H. G., Noma, M., Tekle, A. H., Amazigo, U. V., Diggle, P. J., Giorgi, E. and Remme, J. H. (2014) The geographic distribution of onchocerciasis in the 20 participating countries of the african programme for onchocerciasis control:(2) pre-control endemicity levels and estimated number infected. *Parasites & vectors* **7**(1), 326.

Md Moinuddin

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 327 917 2141
e-mail: moinuddin@stat.unipd.it; moin.unipd@gmail.com

Current Position

Since October 2016; (expected completion: March 2020)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: A Divide and Conquer Approach for Large Spatial Dataset
Supervisor: Prof. Carlo Gaetan
Co-supervisor: Dr. Emanuele Giorgi.

Research interests

- Spatial statistics
- Big-data computation and complex data modeling
- Modeling environmental data
- Epidemiology and disease mapping

Education

July 2009 – June 2010
Master (*laurea specialistica/magistrale***) degree in Statistics**.
Jahangirnagar University, Faculty of Statistics
Title of dissertation: "An advance method of pairwise sequence alignment"
Supervisor: Dr. Mian Arif Shams Adnan
Final mark: A

July 2005 – June 2009
Bachelor degree (*laurea quadriennale***) in Statistics**.
Jahangirnagar University, Faculty of Statistics
Title of dissertation: "Estimating stochastic volatility of stock market using Bata shoe company data"
Supervisor: Prof. Ajit Kumar Majumdar
Final mark: A+.

Visiting periods

April 2019 – June 2019
Lancaster University,
Lancaster, United Kingdom.
Supervisor: Dr. Emanuele Giorgi

Work experience

March 2015 – October 2016
icddr,b.
Statistician .

August 2013 – March 2015
icddr,b.
Research Trainee .

August 2012 – August 2013
icddr,b.
Research Fellow .

Awards and Scholarship

2016-2019
University of Padova PhD scholarship.

Computer skills

- R
- Stata

Language skills

Bengali: native; English: fluent(IELTS: 6.5); Italian: moderate (B1).

Publications

Articles in journals
Gurung R, Jha AK, Pyakurel S, Gurung A, Litorp H, Wrammert J, Jha BK, Paudel P, Rahman SM, Malla H, Sharma S, Gautam M, Linde JE, Moinuddin M, Ewald U, Mlqvist M, Axelin A and Ashish KC ., ... (2019). Scaling Up Safer Birth Bundle Through Quality Improvement in Nepal (SUSTAIN) - a stepped wedge cluster randomized controlled trial in public hospitals. *Implementation Science* **14:65**.

Haider MR, Rahman MM, Moinuddin M, Rahman AE, Ahmed S, Khan MM., ... (2018). Ever-increasing Caesarean section and its economic burden in Bangladesh. *PLoS ONE* **13(12): e0208623**.

Rahman MM, Haider MR, Moinuddin M, Rahman AE, Ahmed S, Khan MM., ... (2018). Determinants of caesarean section in Bangladesh: Cross-sectional analysis of Bangladesh Demographic and Health Survey 2014 Data. *PLoS ONE* **13(9): e0202879**.

Hoque DME, Rahman AE, Perkins L, Islam S, Siddique AB, Moinuddin M, Anwar MR, Mazumder T, Ansar A, Rahman MM, Raihana S, Capello C, Santarelli C, Arifeen SE., ... (2018). Knowledge and involvement of husbands in maternal and newborn health in rural Bangladesh. *BMC Pregnancy and Childbirth* **18:247**.

Ansar A, Rahman AE, Romero L, Haider MR, Rahman MM, Moinuddin M, Siddique MAB, Mamun MA, Mazumder T, Pirani SP, Mathias RG, Arifeen SE, Hoque DME., ... (2018). Systematic review and meta-analysis of global birth prevalence of clubfoot: a study protocol *BMJ Open* **8:e019246**.

Mazumder T, Rahman AE, Hoque E, Mahmud MA, Siddique MAB, Moinuddin M, Rahman MM, Capello C, Arifeen SE, Santarelli C, Perkins J., ... (2018). Advancing People-Centered Maternal and Newborn Health Care through Birth Preparedness and Complication Readiness in Rural Bangladesh *International Journal of Person Centered Medicine,* **7(2)**, 107–117.

Moinuddin M, Christou A, Hoque DME, Tahsina T, Salam SS, Billah SM, et al., ... (2017). Advancing People-Centered Maternal and Newborn Health Care through Birth Preparedness and Complication Readiness in Rural Bangladesh. *PLoS ONE,* **12(12): e0189365**.

Haider MR, Rahman MM, Moinuddin M, Rahman AE, Ahmed S, Khan MM., ... (2017). Impact of maternal and neonatal health initiatives on inequity in maternal health care utilization in Bangladesh. *PLoS ONE,* **12(7): e0181408**.

Rahman AE, Iqbal A, Hoque DME, Moinuddin M, Zaman SB, Rahman QS-u, et al., ... (2017). Managing Neonatal and Early Childhood Syndromic Sepsis in Sub-District Hospitals in Resource Poor Settings: Improvement in Quality of Care through Introduction of a Package of Interventions in Rural Bangladesh. *PLoS ONE,* **2 (1): e0170267**.

Raihana S, Dunsmuir D, Huda TM, Zhou G, Rahman QS, Garde A, Moinuddin M, Karlen W, Dumont GA, Kissoon N, Arifeen SE, Larson C, Ansermino JM., ... (2015). Development and internal validation of a predictive model including pulse oximetry for hospitalization of under-five children in Bangladesh. *PLoS ONE,* **10(11): e0143213**.

Rahman AE, Moinuddin M, Molla M, Worku A, Hurt L, Kirkwood B, Mohan SB, Mazumder S, Bhutta Z, Raza F, Mrema S, Masanja H, Kadobera D, Waiswa P, Bahl R, Zangenberg M, Muhe L, and on behalf of the Persistent Diarrhoea Research Group., ... (2014). Childhood diarrhoeal deaths in seven low- and middle-income countries. *Bulletin of the World Health Organization* **92(9)**,64–71,.

Adnan MAS, Moinuddin M, Roy S, Jaman MR., ... (2011). An alternative method of pair-wise sequence alignment. *In JSM Proceedings, Section on Statistics in Epidemiology. Alexandria, VA: American Statistical Association*, 2942-2951,.

Conference presentations

Moinuddin, M., Gaetan, C., Giorgi, E., (2019). A divide and conquer approach for large spatial dataset. (Oral presentation) *Spatial Statistics 2019*, Sitges, Barcelona, Spain, 10-13 July, 2019.

Moinuddin, M., Gaetan, C., Giorgi, E., (2019). A divide and conquer approach for large spatial dataset. (Poster) *Big data in geoscience workshop 2019*, Lancaster, UK, 19-20 June, 2019.

Teaching experience

November 2019 – February 2020
Fundamentals of Probability and Statistics
Master of Environmental Sciences, Master of Computer Sciences, and Master of Conservation Sciences and Heritage Sciences
Delivering lectures, 30 hours
Cá Foscari University of Venice, Italy

16–20 December 2019
Data Management and Statistical Analysis
PhD, MS and Research assistants
Workshop, 35 hours
International Maternal and Child Health, Uppsala University, Sweden

Other Interests

Traveling
Meeting new people

References

Prof. Carlo Gaetan
Cá Foscari University of Venice
Edificio Zeta, Via Torino, 155
I-30172 Mestre (VE) ITALY
Phone: +39 041 234 8404
e-mail: gaetan@unive.it

Dr. Emanuele Giorgi
Lancaster Medical School, Lancaster University
Furness Building, LA1 4YG, Lancaster, UK
Phone: +44 1524 594319
e-mail: e.giorgi@lancaster.ac.uk