Comparing Models Suggested for Large scale spatial statistics: Case study

Abstract

In this project, we will take a literature review on large scale spatial statistics. In the context of spatio temporal statistics, the given data take spatial dependency or time dependency additionally. Moreover, the format of data could be Geostatistical or Areal (even point process). Figure 1 illustrates this categorization and corresponding implementation of package in R. As denoted by light green in Figure 1, there are many packages in R which model geostatistical spatial data, so we will review those packages. Specifically, we will briefly review underlying theories and compare krigging results by various criteria with satellite data.

1. Introduction

In this section, we will briefly see why large scale spatial data problem matters in spatial statistics. In spatial statistics, we model the underlying spatial stochastic process $\mathbf{w}(\mathbf{s})$ assuming that it follows GP. As data becomes larger and larger, the dimension of $\mathbf{w}(\mathbf{s})$ grows larger and larger, which results in very high dimensional process. Notice that for $\mathbf{w}(\mathbf{s})$ which is $n \times 1$ vector, its covariance is $\mathbf{C}(\cdot, \cdot \mid \boldsymbol{\theta})$ whose dimension is $n \times n$. We assume that covariance matrix of $Y(\mathbf{s})$ is $\boldsymbol{\Sigma} = \mathbf{C} + \tau^2\mathbf{I}$, whose dimension is $n \times n$. [20] shows that in calculating krigging and its standard error, we should compute the inverse of $\boldsymbol{\Sigma}$. Computation complexity of calculating inverse of $n \times n$ matrix is $O(n^3)$, which requires tremendous time to do it. This is the reason why many methodologies were developed to cope with large scale spatial statistics.

2. Model for Point-Referenced Data

Consider q-variate spatial process over $\mathbb{R}^d$. Suppose that $\mathbf{w}(\mathbf{s}) \sim GP(\mathbf{0}, \mathbf{C}(\cdot, \cdot \mid \boldsymbol{\theta}))$, denote a zero-centered q-variate Gaussian process, where $\mathbf{w}(\mathbf{s}) \in \mathbb{R}^q$ for all $\mathbf{s} \in \mathcal{D} \subseteq \mathbb{R}^d$. Let $\mathcal{S} = \mathbf{s}_1, \cdots, \mathbf{s}_k$ be a fixed collection of distinct locations in $\mathcal{D}$, which we call the reference set. We are interested in spatial regression model such as

$$Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})^T\boldsymbol{\beta} + w(\mathbf{s}) + \epsilon(\mathbf{s}) \tag{1}$$

where $w(\mathbf{s})$ denotes spatially correlated process, often called spatial random effects and $\epsilon(\mathbf{s})$ is an white noise process. They are typically assumed to be independent each other. Also, $\mathbf{C}$ is assumed as a matern covariance matrix and $\epsilon(\mathbf{s})$ follows iid normal. That is, denoting $e(\mathbf{s}) = w(\mathbf{s}) + \epsilon(\mathbf{s})$, $e(\mathbf{s}) \sim \mathcal{N}(0, \mathbf{C} + \tau^2\mathbf{I})$.
Customary bayesian hierarchical models are constructed as

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2) \times N(\mathbf{w} \mid \mathbf{0}, \mathbf{C}) \times N(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta} + \mathbf{w}, \tau^2\mathbf{I}) \tag{2}$$

3. Brief Reviews of Theories

Hierarchical Nearest Neighbor Gaussian Process [spNNGP package]

Abhirup Datta et al. [1] suggested alternating standard gaussian process with Nearest Neighbor Gaussian Process(NNGP).
The joint distribution of $p(\mathbf{w}_\mathcal{S})$ can be decomposed as

$$\begin{aligned} p(\mathbf{w}_\mathcal{S}) = p(\mathbf{w}(\mathbf{s_1}))\, p(\mathbf{w}(\mathbf{s_2}) \mid \mathbf{w}(\mathbf{s_1})) \\ \cdots p(\mathbf{w}(\mathbf{s_k}) \mid \mathbf{w}(\mathbf{s_{k-1}}), \ldots, \mathbf{w}(\mathbf{s_1}) \end{aligned} \tag{3}$$

For every $\mathbf{s}_i \in \mathcal{S}$, let a smaller conditioning set $N(\mathbf{s}_i) \subset \mathcal{S} \setminus \mathbf{s}_i$, whose size is at most $m << k$. Main idea of NNP is that to approximate $p(\mathbf{w}_\mathcal{S})$ as

$$\tilde{p}(\mathbf{w}_\mathcal{S}) = \prod_{i=1}^{k} p(\mathbf{w}(\mathbf{s_i}) \mid \mathbf{w}_{N(\mathbf{s}_i)}) \tag{4}$$

Therefore, the NNGP prior implies that the spatially correlated vector $\mathbf{w}$ now follows a gaussian process with different covariance matrix, i.e., $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \tilde{\mathbf{C}}(\boldsymbol{\theta}))$. The covariance

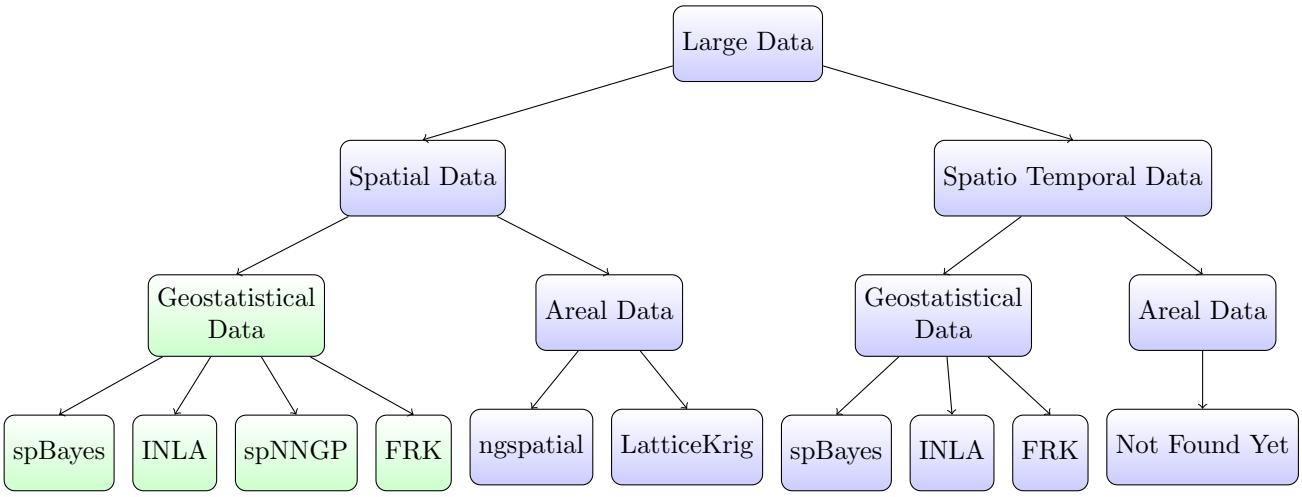

Figure 1: Diagram of data format and corresponding R package

matrix of NNGP $\tilde{\mathbf{C}}$ ensures the sparsity compared with $\mathbf{C}$, which is only $O(n)$ in computation [1]. The gibbs sampler from conjugate and metropolis random-walk step are proposed in [1]. Especially, $\mathbf{w}_{s_i}$ is updated sequentially through full conditional $\mathbf{w}_{s_i} \mid \cdot \sim \mathcal{N}$. However, this sequentially updating of elements could slow the convergence, because updating a high dimensional latent random effect vector is prone to high autocorrelation. As an alternative and extension of [1], [2] directly derive marginal gaussian process for the response, i.e., $\{y(\mathbf{s})\}$. Therefore the response model is

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\beta, \tilde{\mathbf{\Sigma}}) \tag{5}$$

Using this response model, the MCMC-Free exact bayesian inference using conjugate NNGP is proposed [2]. Combining with priors of $\boldsymbol{\beta}, \sigma^2$ leads to conjugate Normal-Inverse-Gamma posterior distribution. Instead of sampling from the posterior directly, it provides an algorithm which circumvents MCMC based iterative sampling but at the same time, its accuracy is quite comparable with original MCMC and computation time is only $O(n)$. This model is called Conjugate NNGP which is implemented in spNNGP package in R [4]

Predictive Process Model [spBayes package]

Consider a set of 'knots', $\mathcal{S}^* = \{\mathbf{s}_1^*, \cdots, \mathbf{s}_m^*\}$ which could form a subset of entire observed locations $\mathcal{S}$. The gaussian process using these m locations is $\mathbf{w}^* = [w(\mathbf{s}_i^*)]_{i=1}^m \sim MVN\{\mathbf{0}, C^*(\boldsymbol{\theta})\}$, where $C^*(\boldsymbol{\theta})$ is the corresponding $m \times m$ covariance matrix. The spatial krigging at $\mathbf{s}_0$ is expressed as $\tilde{w}(\mathbf{s}_0) = E[w(\mathbf{s})_0 \mid \mathbf{w}^*] = \mathbf{c}^T(\mathbf{s}_0; \boldsymbol{\theta}) C^{*-1}(\boldsymbol{\theta}) \mathbf{w}^*$, where $\mathbf{c}^T(\mathbf{s}_0; \boldsymbol{\theta}) = [C(\mathbf{s}_0, \mathbf{s}_j^*; \boldsymbol{\theta})]_{j=1}^m$, cross covariance vector. This single-site interpolator defines a spatial process $\tilde{w}(\mathbf{s}) \sim GP\{0, \tilde{C}(\cdot)\}$. We refer to this spatial process of single-site interpolator as the *predictive process* [5] derived from the parent process $w(\mathbf{s})$. Notice that

$$\tilde{C}(\mathbf{s}, \mathbf{s}^{'}; \boldsymbol{\theta}) = \mathbf{c}^T(\mathbf{s}; \boldsymbol{\theta}) C^{*-1} \mathbf{c}(\mathbf{s}^{'}, \boldsymbol{\theta}) \tag{6}$$

So, this spatial process is non-stationary whereas the standard spatial model assume stationary process. Replacing $w(\mathbf{s})$ with non-stationary process $\tilde{w}(\mathbf{s})$, we get the following predictive process model

$$Y(\mathbf{s}) = \mathbf{x}^T \boldsymbol{\beta} + \tilde{w}(\mathbf{s}) + \epsilon(\mathbf{s}) \tag{7}$$

Note that because we choose m knots, the predictive process is m dimensional. So we can work with $m \times m$ covariance matrix from which we can see dimension reduction immediately.

In smoothing spline, it is natural to place knots at every data point, i.e., natural cubic spline [8]. However, in spatial literature, there are lots of points, so placing knots at every data point is almost impossible. Many methods were suggested regarding to selection of knots [9, 10, 11]. However, the predictive process method still suffers from the choice of knot, which is the same problem in smoothing literature.

Stochastic Partial Differential Equation [INLA package]

[17] suggests a real breakthrough which considers a stochastic partial differential equation (SPDE) whose solution is a Gaussian field with

matern covariance, i.e., GMRF. GMRF has been used popularly in spatio-temporal statistics because the matern covariance is considered a standard covariance structure which assumes the stationary and isotropic. Despite its popularity, it has some drawbacks, which were stated earlier as *big n problem*. However, [17] propose to represent a GF with matern covariance with the solution of SPDE. More specifically, [17] uses the Finite Element Method (FEM) to provide a solution with a SPDE with which GMRF is expressed. It develops this solution by considering basis functions using the traingular mesh nodes. Letting $w(\mathbf{s})$ following GMRF, we write

$$w(\mathbf{s}) = \sum_{k=1}^{m} \phi_k(\mathbf{s}) w_k^{\star} \tag{8}$$

Using this linear combination of basis function, we write the spatial model as

$$Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})^T \boldsymbol{\beta} + \boldsymbol{\phi}\boldsymbol{w}^{\star}(\mathbf{s}) + \epsilon(\mathbf{s}) = \boldsymbol{A}z(\mathbf{s}) + \epsilon(\mathbf{s}) \tag{9}$$

where $\mathbf{A} = (\boldsymbol{\phi}, \mathbf{X})$.
Note that the transformed model (9) is a subclass of structured additive regression models which are widely used to perform approximate Bayesian inference. In these models, the response variable y_i is assumed to belong to an exponential family whose mean μ_i is associated with link function $g(\cdot)$, as $g(\mu_i) = \eta_i$. In GLM literature, η_i is linear function with respect to other covariates. In structured addtive regression models, η_i is combinations of other effects, say,

$$\eta_i = \mu + \sum_j \beta_j z_{ij} + \sum_k w_k f^k(u_{ik}) \tag{10}$$

If we assign gaussian prior on $\mu, \boldsymbol{\beta}, \mathbf{f}^k$, we call it as *latent Gaussian models*.
Let's define some notation to simplify following discussion. Let $\mathbf{x}$ be all the n gaussian variables $\{\eta_i\}, \alpha, \{f^{(j)}\}, \{\beta_k\}$ and $\boldsymbol{\theta}$ be all hyperparameters. Foramlly, the model can be written as

$$\begin{aligned} \mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta} &\sim \prod \pi(y_i \mid \eta_i, \boldsymbol{\theta}) \\ \mathbf{x} \mid \boldsymbol{\theta} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}^{-1}(\boldsymbol{\theta})) \\ \boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}) \end{aligned} \tag{11}$$

In INLA [12], $\mathbf{x} \mid \boldsymbol{\theta}$ follows Gaussian Markov Random Field (GMRF) [18]. By the property of GMRF, the precision matrix $\boldsymbol{Q}(\boldsymbol{\theta})$ is sparse matrix, which means $\mathbf{x}_i \perp \mathbf{x}_j \mid \mathbf{x}_{ij} \iff Q_{ij} = 0$.

The posterior then follows,

$$\begin{aligned} \pi(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y}) &\propto \pi(\boldsymbol{\theta}) |\boldsymbol{Q}(\boldsymbol{\theta})|^{1/2} \\ &\quad \times exp[-\frac{1}{2}\mathbf{x}^T \boldsymbol{Q}(\boldsymbol{\theta})\mathbf{x} + \sum log\{\pi(y_i \mid x_i, \boldsymbol{\theta})\}] \end{aligned} \tag{12}$$

Often the main interest lies in the marginal posterior of the latent field $\pi(x_i \mid \mathbf{y})$ or the hyperparameters $\pi(\theta_j \mid \mathbf{y})$. To get these, [19] used gaussian approximations, which have some problems. INLA tries to solve this through nested approximations.
The distributions of main interest are

$$\begin{aligned} \pi(\theta_j \mid \mathbf{y}) &= \int \pi(\boldsymbol{\theta} \mid \mathbf{y})\, d\boldsymbol{\theta}_j \\ \pi(x_i \mid \mathbf{y}) &= \int \pi(x_i \mid \boldsymbol{\theta}, \mathbf{y}) \pi(\boldsymbol{\theta} \mid \mathbf{y})\, d\boldsymbol{\theta} \end{aligned} \tag{13}$$

Therefore, depending on which method we use to approximate $\pi(\boldsymbol{\theta} \mid \mathbf{y})$, $\pi(x_i \mid \boldsymbol{\theta}, \mathbf{y})$, we can reduce computation regarding integration. INLA approximates these quantities as

$$\begin{aligned} \tilde{\pi}(\boldsymbol{\theta}^k \mid \mathbf{y}) &\propto \frac{\pi(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})^k \pi(\mathbf{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{\pi(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})} \\ \tilde{\pi}(x_i \mid \mathbf{y}) &= \sum_{k=1}^{K} \tilde{\pi}(x_i \mid \boldsymbol{\theta}^k, \mathbf{y}) \tilde{\pi}(\boldsymbol{\theta}^k \mid \mathbf{y}) \Delta_k \end{aligned} \tag{14}$$

where $\boldsymbol{\theta}^k$ is a specific value of the hyperparameters vector, $\tilde{\pi}_G$ is a gaussian approximation to the full conditional $\mathbf{x} \mid \boldsymbol{\theta}^k, \mathbf{y}$, Δ_k are appropriate weights.
The name *nested approximations* in INLA comes from the procedure that for $k = 1, \cdots, K$ once we compute $\boldsymbol{\theta}^k$ (approximated values) and $\tilde{\pi}(x_i \mid \boldsymbol{\theta}^k, \mathbf{y})$ (approximated by one of Laplce, Simplified Laplace or Gaussian), we get $\tilde{\pi}(x_i \mid \mathbf{y})$ via the numerical integration as denoted in (10).

Fixed Rank Krigging [FRK package]

Recall that the computational cost of $O(n^3)$ is the result of inverting covariance matrix $\mathbf{C}$, which is essentially done when krigging. However, [20] shows that with modification from stationary covariance matrix to non-stationary covariance matrix, we can reduce computations upto $O(n)$ which is dramatic reduction. This is done using fixed r number of basis functions and modeling original covariance matrix using theses basis functions. Specifically, if we choose a set of r basis functions, $\mathbf{S}(\mathbf{u}) \equiv (S_1(\mathbf{u}), \cdots, S_r(\mathbf{u}))$, where $\mathbf{u} \in \mathbb{R}$, we model $C(\mathbf{u}, \mathbf{v}) = \mathbf{S}(\mathbf{u})^{'} \mathbf{K} \mathbf{S}(\mathbf{v})$, for any $r \times r$ positive

definite matrix $\mathbf{K}$. This can ve shown to be a non-negative-definite function hence is a valid covariance function. Using this expression, we can notice that

$$\mathbf{\Sigma} = \mathbf{S}\mathbf{K}\mathbf{S}^{'} + \tau^2\mathbf{I} \tag{15}$$

The problem of using Matern covariance matrix in gaussian matrix is calculating $\mathbf{\Sigma}^{-1}$ requires $O(n^3)$ computation. However, from [20] we can show that

$$\mathbf{\Sigma} = (\sigma^2)^{-1} - (\sigma^2)^{-1}\mathbf{S}\{\mathbf{K}^{-1} + \mathbf{S}^{'}(\sigma^2)^{-1}\mathbf{S}\}^{-1}\mathbf{S}^{'}(\sigma^2)^{-1} \tag{16}$$

which involves getting inverse of $r \times r$ positive definite matrices. When calculating the krigging predictor and its standard error, same logic can be applied, where computation is linear n.

We are now left with the selection of basis function. Because there is no assumption with orthogonal basis function, the choice of $\mathbf{S}(\cdot)$ is unrestricted and may include various basis functions. Among them, [20] recommends to use basis functions that are multiresolutional because multiresolutional components of $\mathbf{S}(\cdot)$ allow many spatial scales of variation to be captured.

4. Experiment Results

In this section, we will compare krigging results, running time of each spatial model. We will first elaborate which data we use and various criteria to compare krigging results.

Data

The data used in this experiment consist of daytime land surface temperature measured by MODIS satellite on August 4, 2016. It was used used in [22] for the same purpose. Among many regions and date observed by satellite, this part of the data have a little missing values. That is, 1.1% of total data were corrupted by cloud which results in 148,309 observations of temperature out of 150,000. Because the observations were on August 4, 2016, they have no time dependency but spatial dependency. Therefore, we can say that it is large scale spatial data.

For this data, the response variable is *temperature* and there are no covariates, so the linear trend only consists of intercept term. In other words, the basic spatial model for data is

$$Y(\mathbf{s}) = \beta_0\mathbf{1} + w(\mathbf{s}) + \epsilon(\mathbf{s}) \tag{17}$$

Criteria to Evaluate

We evaluate the various models through 1) RMSE 2) MAE (Mean Abolute Error) 3) CRPS (Continuous Rank Probability Score). We briefly mention CRPS [23]. Following the guideline in [22], we calculate $(U - L)/(2 \times \Phi^{-1}(0.975))$ where $\Phi(\cdot)$ denote cdf of standard normal distribution and take it as predictive standard error. Assuming the associated predictive distribution was well approximated by a gaussian distribution, we calculate the CRPS by setting zero mean and standard deviation equal to predictive standard error calculated above.

Experiment Details

When using NNGP as a prior for $\mathbf{w}$, we use its improved version [2], implemented as `spConjNNGP` function in snngp package in R [4]. As detailed in [2], Conjugate NNGP estimates $\phi, \alpha = \tau^2/\sigma^2$ using K-fold cross-validation, which we follow in this experiment too.

When using predictive process model, we use the function `spLM` in [7]. Note that it is important to choose appropriate knots, which is often a challenge to do. In [5], the knots were chosen as lattice grid partitioned regularly. In addition, following desings in [6] it chooses more knots by lattice plus close pair configuration and lattice plus infill configuration. In this experiment, we will just use 14×14 regularly spotted knots, as denoted by Figure 2 (a). Unfortunately, we do not show the results of predictive process model because we could not get the krigging of test data although we wait for almost a day. Since in practice, one day for krigging is already quite long time to wait for the results, we judge that it is meaningless to wait longer.

Next, we approach the large scale data by SPDE and solve it using INLA package in R. We follow the tutorial and use some advanced options referencing [14, 15, 16]. The procedure of INLA is more technical than other approaches. First of all, we should build the triangulated mesh on top of which the SPDE/GMRF representation is to be built. INLA package provides vest choices of mesh and we choose commonly used mesh option, that is `inla.mesh.2d`. We make mesh with training data. Next we define the SPDE model with matern. Then we make basis matrix, denoted by A in [12]. Finally, we put some fixed effects, random effects all together which is often called as stack. After this step, we

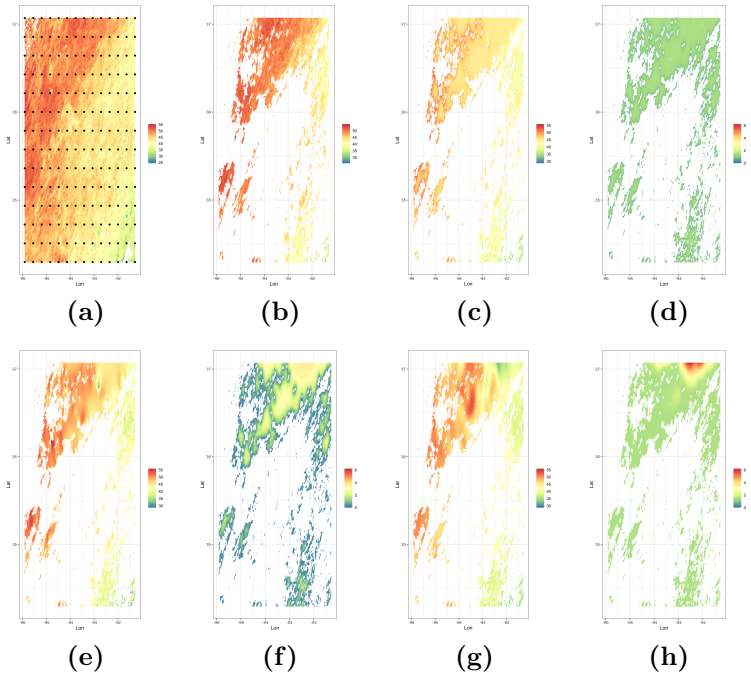

Figure 2: **(a)** Temperature in whole region with knots used in PPM **(b)** Temperature in test data region **(c)** Krigged temperature by snngp **(d)** S.E of (c) **(e)** Krigged temeprature by INLA **(f)** S.E of (e) **(g)** Krigged temeprature by FRK **(h)** S.E of (g)

finally set up INLA model.
When modeling using FRK, we follow the steps elaborated in [21]. We assign Basic Areal Units (BAUs) manually using whole data and generate 569 basis functions on the plane. Then, we krig the test data and calculate the s.e.

	MSE	MAE	CRPS	Time (min)
sNNGP	2.49	1.9	1.46	28
INLA	2.01	1.47	1.15	40
FRK	3.26	2.38	1.77	1.5

Table 1: Summary of results

Results

In Figure 2, the whole data are plotted in (a), together with knots used when predictive process modeling, only test data are plotted in (b) and (c) $\sim$ (h) denote the krigging results and estimated s.e. by each method. We set the scale of krigging plots and s.e. plots identically for better comparing each other. For all methods, the true value is underestimated. The overall estimated s.e. of snngp is very low and that of INLA is relatively high. To compare more precisely, we compute the RMSE, MAE, CRPS as mentioned before.

Note that for all criteria, INLA performs best among three methods. We can guess this result from Figure 2 (c), (e), (g) in that (c) is the most similar with the actual plot (b). Note that although overall s.e. of snngp is relatively lower than other methods, the calculated CRPS of INLA method is the smalles one. However, the computation time (training + krigging) is larger than other methods. In particular, the computation time of FRK is notably smaller than others. In conclusion, among three methods, although INLA takes the longest time to complete modeling and krigging, it performs better than other methods in terms of performance criteria.

5. Conclusion

In this project, we compare methodologies developed for large scare spatial statistics. If studied deeper about the architecture so that setting the hyperparameters or priors properly, the results of each models may improve in terms of performance and computation time.

References

[1] Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets, Abhirup Datta et al., 2015.

[2] Efficient Algorithms for Bayesian Nearest Neighbor Gaussian Processes, Finley et al., 2019.

[3] spNNGP R package for Nearest Neighbor Gaussian Process models, Finley et al., 2019.

[4] package 'spNNGP', Finley et al., 2020.

[5] Gaussian predictive process models for large spatial data sets, Sudipto Banerjee et al., 2008.

[6] Bayesian Geostatistical Design, Diggle et al., 2006.

[7] Package 'spBayes', Finley et al., 2020.

[8] Functional Data Analysis, Ramsay et al., 2005.

[9] Design of air-quality monitoring networks, Nychka et al., 2007.

[10] Spatially balanced sampling of natural resources, Stevens et al., 2004.

[11] Bayesian geostatistical design, Diggle et al., 2006.

[12] Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, Rue et al., 2009.

[13] Integrated Nested Laplace Approximations (INLA), Martino et al., 2019.

[14] Bayesian Spatial Modelling with R-INLA, Lindgren et al., 2015.

[15] Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA, Elias T. et al., 2019.

[16] Bayesian inference with INLA, Virgilio et al., 2020.

[17] An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach, Lindgren et al., 2011.

[18] Gaussian Markov random fields: theory and applications, Rue et al., 2005.

[19] Approximate Bayesian inference for hierarchical Gaussian Markov random fields models, Rue et al, 2007.

[20] Fixed rank kriging for very large spatial data sets, Cressie et al., 2008.

[21] FRK: An R Package for Spatial and Spatio-Temporal Prediction with Large Datasets, Andrew et al., 2018.

[22] A Case Study Competition Among Methods for Analyzing Large Spatial Data, Heaton et al., 2018.

[23] Strictly Proper Scoring Rules, Prediction, and Estimation, Gneiting et al., 2007