SAND2023-05368T

ADVANCEMENTS IN ENVIRONMENTAL STATISTICS CONCERNING MULTIPLE DATA SOURCES

BY

MAURICIO CAMPOS

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois Urbana-Champaign, 2023

Urbana, Illinois

Doctoral Committee:

Professor Bo Li, Chair
Professor Douglas G. Simpson
Clinical Associate Professor Trevor H. Park
Clinical Associate Professor Lelys Bravo de Guenni

Abstract

This work considers three different applications of environmental spatial statistics. Two of them relate to integrating data coming from different sources, while the other does the opposite and instead breaks down an aggregated response into several components of interest. The contributions of this work are separated by application into three parts.

The first part is motivated by the interest in studying the biotic responses of species during the Last Glacial Maximum (LGM) due to rapid anthropogenic climate change. During this period, species retreated to highly spatially restricted geographic regions where survival was possible, known as glacial micro-refugia, from which they migrated and expanded when conditions became more suitable. Several distinct sources of evidence have contributed to developing a new understanding of how these regions might have impacted the sustainability of the natural populations of many species. Pollen records in Eastern Beringia (EB) have been used to explore the possibility that the region harbored glacial refugia for several plants from the arctic tundra and/or the boreal forest biomes common to the region. Our study focuses on *Alnus viridis* and *Picea glauca*, two predominant species of arcto-boreal vegetation. We propose to integrate genomic, SDM, and existing fossil data in a hierarchical Bayesian modeling (HBM) framework to determine whether multiple refugia existed in isolated geographic areas. This study demonstrates how the flexibility of HBMs makes the formal synthesis of such disparate data sources feasible. Our results highlight the regions of plausible refugia that can guide future investigations into studying the role of glacial refugia during climate change.

The second part reverses the data integration of the first in hopes of utilizing present technologies better for the purposes of crop monitoring. The amount of carbon assimilated by plants through photosynthesis, called Gross Primary Productivity (GPP), is the largest carbon flux between the

terrestrial biosphere and the atmosphere and, if quantified accurately, can grant insights into understanding several ecosystem functions as well as the impact of climate change to crop yields, in particular corn and soy. Recently, satellite-based measurements of solar-induced chlorophyll fluorescence (SIF) have been used as a strong proxy to measure GPP. SIF values will depend on the type of vegetation land cover; thus, the observed values can be decomposed into the specific vegetation type components to obtain its particular SIF yield information. We propose to implement a spatially varying coefficient regression model where the coefficients represent the specific SIF yields. For each land type coefficient, we induce spatial smoothness by penalizing the square deviations among adjacent sites according to some data-driven threshold value. The adjacent sites are chosen according to a minimum spanning tree (MST) in order to reduce redundancy in site pairing. Special characteristics of the data impose additional challenges, such as a non-negativity constraint on the estimations, as well as the presence of deterministic information that changes the structure of the MST. This study is able to retrieve accurate and fast results for the two main crops of interest when compared to other similar methods.

Finally, the third part returns to data integration in handling the change of support problem in spatial statistics. Environmental applications are highly dependent on accurate and complete datasets of world temperature. However, the recollection of these datasets is limited by technology and poses additional challenges that must be handled before doing any environmental analysis. In particular, the change of support problem occurs when trying to combine data that has been collected at different resolutions. Some data is collected in situ, whereas other can be collected by satellites that cover wider areas. This study uses INLA to accurately handle data coming from the Integrated Global Radiosonde Archive (IGRA) and from the TIROS Operational Vertical Sounder (TOVS), both from the National Oceanic and Atmospheric Administration (NOAA), from 1990 to 1993. Both datasets are considered to measure the same latent process but in different ways that must be integrated to produce a complete picture of global temperature. This must also be done close to real-time, so an alternative model is also proposed to speed computations at the cost of accuracy. We are able to obtain similar results from both approaches as well as provide accurate uncertainty measurements for both. These results can then be used for future applications of environmental studies.

To my parents. Thank you for everything.

Acknowledgments

I want to first thank the teachers who made this possible. In particular, I want to thank my advisor, Bo Li, who contributed tremendous time and effort to guide me throughout this whole process. I am fortunate to have had Bo push me to become a better statistician in this time.

I would also like to thank all the collaborators with whom I've worked in all my different project. I thank Guillaume de Lafontaine, Joe Napier and Feng Sheng Hu in helping me understand the details about different paleoecological sources of evidence. I thank Kaiyu Guan and his team in providing several references that helped understand the relation between SIF and different vegetation types. I also want to thank Audrey McCombs, Justin Li, Gabriel Huerta and Lyndsay Shand from Sandia National Laboratories for welcoming me into their time and give me a chance to work alongside them.

I also thank my parents, Raul and Liliana, for always encouraging and supporting me, and for being great parents overall. I also thank the friends that have stuck with me since the beginning and those that I have made along the way, both in Urbana-Champaign and Costa Rica, for their support. I wouldn't have made it this far without the good times we had.

Table of contents

Chapter 1

Data Fusion of Temperature Datasets Using INLA

1.1 Introduction

A very classic problem in spatial statistics is combining information that is gathered at differing spatial resolutions. this happens due to the advances in remote sensing and satellite imagery which allows for data to be gathered at different resolutions. This is normally referred to as 'areal' data, since the information retrieved is aggregated over a particular area given by the resolution of the measuring instrument.

Opposed to this is another classical type of spatial data, referred to as geostatistical data. This is considered a spatial process indexed over a continuous space. This information is typically gathered by instruments that are fixed in a particular known location for a given time, like weather stations, weather balloons, etc. For the remainder of this work we will refer to this data as 'point' data, in contrast to areal data. Nonetheless, this does not mean that we are talking about point data in the classical spatial statistics sense, where the location is random.

When there is a mismatch or inconsistency between the spatial units at which data are observed or measured and the spatial units at which the analysis or inference is desired, we are presented with what is known as the change of support problem. This problem is particularly relevant in the field of geostatistics, where spatial data are commonly collected at one set of locations or spatial

resolutions but are needed or desired at a different set of locations or spatial resolutions for modeling, prediction, or decision-making purposes.

One particular case of the change of support problem that is very common to environmental sciences is trying to combine information from different resolutions. Typically, we are interested in using both areal and point data to estimate a process at a smaller resolution. This is the main goal of this project, where we try to combine the areal data concerning surface temperatures obtained from several instruments on board of NOAA satellites and the point data from a collection of historical and near-real-time radiosonde and pilot balloon observations.

The main goal of combining these data sets is to produce a more complete temperature map of the entire world. The point data is mostly in land and in the northern hemisphere. Areal data is collected throughout the globe, but with a great quantity of missing areas due to cloud coverage and other issues that arise from satellite imaging. This leaves an incomplete picture of global temperature that we hope to complete by combining both data sets in a modeling framework.

This work is part of a bigger project that seeks to create a complete data set of world surface temperature from 1990-1993 for use in other research projects. Typically, reanalysis data is used to fulfill this need, but it either lacks uncertainty calculations or they take too much time to compute. The goal of this project is to find a method that is able to interpolate the observed data to the the entire globe in a time-efficient manner and that also takes into account the differences in handling areal and point data. This project intends to do that by looking at how INLA is able to treat areal and point data differently.

The following project is divided into several sections. Section 1.2 introduces where the observed areal and point data comes from, as well as the sampling limitations that are present in them that challenge inference. Section 1.3 discusses how areal and point data are modeled differently and how INLA respects this distinction, as well as an alternative hierarchical model approach that is faster to fit but loses accuracy. Section 1.4 shows the results of several simulations to study how different sampling scenarios affect the predictions, as well as how much the prediction is affected by using the alternative model. Finally, results are shown in Section 1.5 with a discussion made in Section 1.6.

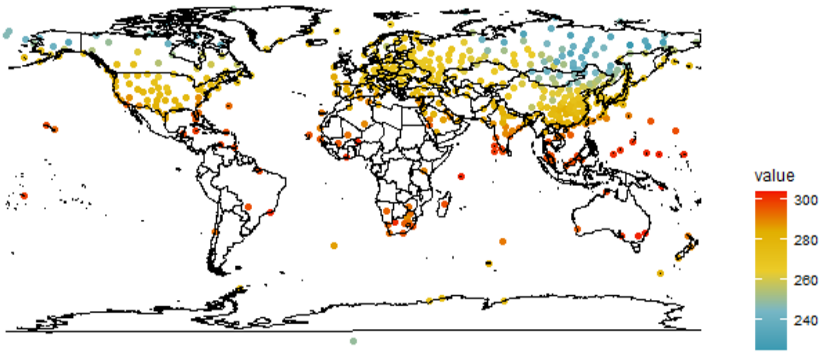


Figure 1.1: Point temperature data obtained from NOAA's Integrated Global Radiosonode Archive (IGRA) for January 1st, 1990.

1.2 Data

Point and areal world surface temperature data are collected daily from 1990 to 1993. For this particular project, we will be mostly focusing on January 1st, 1990. Point data is obtained from the Integrated Global Radiosonde Archive (IGRA), which is provided by the National Oceanic and Atmospheric Administration (NOAA) (Durre et al. 2016). This data consists of radiosonde and pilot balloon observations from more than 2,800 globally distributed stations that date back to 1905. As seen in one example for January 1st, 1990 in Figure 1.1, most of the data is located in land and in the northern hemisphere. This holds true for the remaining days in the study, which posses an interesting question in how well the data will mix in those regions where we have less or no observations.

Areal data is recovered from NOAA's TIROS Operational Vertical Sounder (TOVS), which is a suite of three instruments that were flown on the NOAA-6 through NOAA-14 Polar-orbiting Operational Environmental Satellites. These instruments were primarily designed for atmospheric sounding are sensitive to surface temperatures (Anyamba and Susskind 1998). Data is recovered in blocks of 1° by 1°, meaning that it creates a grid of size 180×360.

Figure 1.2 displays the areal temperature data retrieved by TOVS. The gray blocks are those were cloud cover made it impossible for the instruments to obtain a reasonable temperature reading. For this particular day, the amount of missing data represented 52.28% of the grid. However, this percentage is variable for other days and is not always as high. Nonetheless, this is the main reason why we must find a way to combine this data with point data to produce a more complete snapshot

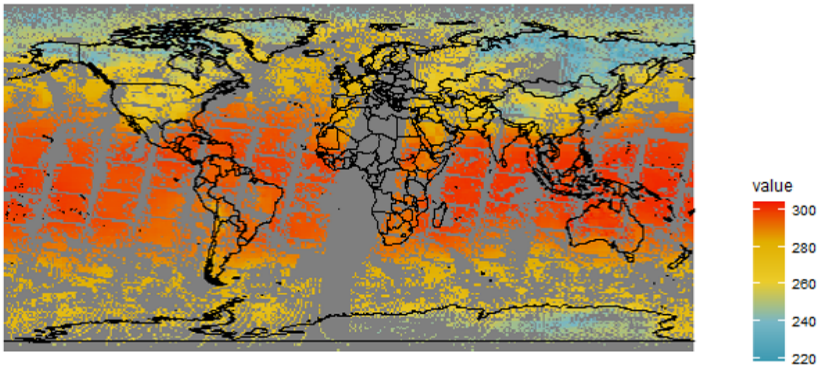


Figure 1.2: Areal temperature data obtained from NOAA's TIROS Operational Vertical Sounder (TOVS) for January 1st, 1990. The gray areas correspond to missing data due to cloud cover.

of daily surface temperature.

1.3 Models and Estimation

We suppose we are studying an underlying spatial process from which we observe continuous observations with some measuring error. Let the underlying spatial process be $S(\boldsymbol{x})$, $\boldsymbol{x} \in D \subset \mathbb{R}^2$ where we assume that $S(\boldsymbol{x})$ is a mean-zero Gaussian process with some covariance function. For a finite set of sites, say $\boldsymbol{x}_i \in D$, $i = 1, 2, ..., n_p$, we model the point data as:

$$y(\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i) + S(\boldsymbol{x}_i) + \varepsilon(\boldsymbol{x}_i) \tag{1.1}$$

Areal data is defined as averages in a block $A_j \subset D$, $j = 1, 2, ..., n_a$, such that

$$y(A_j) = \frac{1}{|A_j|} \int_{A_j} (\mu(\boldsymbol{x}) + S(\boldsymbol{x}))dx \tag{1.2}$$

where $|A_j| > 0$ denotes the area of block A_j. For this project, we will assume that all blocks are observed on a regular grid, and thus have the same shape and area.

1.3.1 Estimation with INLA

The models are fit using the INLA approach (Rue, Martino, and Chopin 2009) with the SPDE approach (Lindgren, Rue, and Lindström 2011), following the handling of the change of support

problem portrayed in Moraga et al. 2017. All of this can be easily applied using the R package R-INLA (Lindgren and Rue 2015).

As mentioned in Chapter **??**, INLA uses a combination of analytical approximations and numerical integration to do Bayesian inference in models that have a latent Gaussian process. Furthermore, using the SPDE approach, modeling can be done in continuous space, but the inference uses the sparse precision matrices of the GMRF defined on the triangular mesh of the spatial domain. The representation of the continuously indexed field, $S(\boldsymbol{x})$, through the discretely indexed GMRF is done by means of a finite basis function defined on the mesh:

$$S(\boldsymbol{x}) = \sum_{m=1}^{M} \phi_m(\boldsymbol{x}) S_m \tag{1.3}$$

where S_m are mean-zero Gaussian weights, $\phi_m(\cdot)$ denotes a piecewise polynomial basis function on each triangle, and M is the number of vertices in the mesh.

The way R-INLA approximates $S(\boldsymbol{x})$ for point data is by calculating the weighted mean of the GMRF estimates in the vertices of the triangle that includes $\boldsymbol{x}$. The weights are given by the barycentric coordinates, i.e. they are proportional to the areas of each of the three subtriangles defined by the point $\boldsymbol{x}$ and the vertices of the triangles. Areal data, $S(A)$ is estimated by taking the average of all the GMRF values in block A. This makes it necessary for each areal block to include at least one vertex of the mesh.

Since INLA uses the SPDE approach, this means that the only stationary covariance function available to fit models is the Matérn covariance function, as defined in Chapter 1. However, this approach allows for a more flexible class of non-stationary models. The details of this can be found in Lindgren, Rue, and Lindström 2011, which explains how it is possible to define the non-stationarity in the SPDE instead of in the covariance function itself. For this model we allow the variance to vary with latitude, such that the variance increases in a polynomial way as we move farther away from the equator.

Another advantage of using the SPDE approach is the gain in computation time by using a sparse precision matrix but also thanks to the dimension reduction obtained by only having to do the estimation in the triangle vertices. Normally, we would expect to have fewer nodes in the triangulation than data. However, this is not always the case when working with areal data. Since

we must consider that each block must have at least one vertex in the triangulation, this means that there is actually a dimension increase when fitting the GMRF. This is typically not an issue as we are still working with sparse precision matrices, unless the resolution is very large (i.e. small areal block) in which case computation time will be affected.

1.3.2 Alternative model

To handle the larger computation times encountered by having to create a very dense mesh we have to handle the model differently to be able to use a sparse mesh. A very simple alternative is to just use the areal data as if it were point data, but obviously this doesn't account for the change of support. This would mean that the uncertainty specific to areal data would be underestimated.

One way to respect the different uncertainties for point and areal data but also create a computationally faster method is to suppose a hierarchical model structure where both areal and point data are treated as observations from a common latent field, but with different uncertainty structures.

The first level consists of modeling both data sets as realizations of the same common latent field, $\mu(x) + S(\boldsymbol{x})$, but with independent and different error structures:

$$\begin{aligned} Y_p(\boldsymbol{x}) &= \mu(\boldsymbol{x}) + S(\boldsymbol{x}) + \epsilon_p, \quad \epsilon_p \sim N(0, \sigma_p^2), \\ Y_a(\boldsymbol{x}) &= \mu(\boldsymbol{x}) + S(\boldsymbol{x}) + \epsilon_a, \quad \epsilon_a \sim N(0, \sigma_a^2), \end{aligned} \tag{1.4}$$

In comparison to equations (1.1) and (1.2), point data is modeled the same way but areal data is now modeled as point data but with a different error structure that hopefully is able to capture the different uncertainty that is characteristic of this type of data.

For simplicity, we will model $\mu(\boldsymbol{x})$ as a constant μ, but it could be interesting in the future to include some more information that depends on $\boldsymbol{x}$. However, as for now, we do not have any more insight into how this mean structure could look like.

The second level of this hierarchical model describes the latent field, $S(\boldsymbol{x})$, as mean zero Gaussian process with some covariance function:

$$S(\boldsymbol{x}) \sim GP(0, \Sigma) \tag{1.5}$$

The covariance function is modeled using a fixed smoothness parameter, $\nu = 1$, and two unknown parameters for the spatially varying variance, $\sigma^2(\boldsymbol{x})$, and stationary range parameter, ρ. The variance is modeled as $\log(\sigma(\boldsymbol{x})) = \theta_1 + \theta_2 \cdot \text{scale}(x_2)^2$, where x_2 refers to latitude and the scale$(\cdot)$ function centers the latitude around it's mean and scales it to have a variance of 1.

Finally, the third level of the model would be given by the prior distributions on the parameters:

$$\begin{aligned} \mu &\sim N(0, 1000), \\ \log(1/\sigma_p^2), log(1/\sigma_a^2) &\sim \text{LogGamma}(1, 0.00005), \\ (\theta_1, \theta_2, \log\rho)^T &\sim N\left((0, 0, \log(142))^T, 100 \cdot \boldsymbol{I}_3\right). \end{aligned}$$

The log range parameter of the latent field is centered at $\log(142)$ since it represents about a third of the maximum distance of the spatial domain. Nonetheless, the variance is kept high to still be considered a mostly uninformative prior like all the others used.

1.4 Simulation Study

The simulation study is divided into two sections. The first one plans to evaluate how well INLA is able to combine point and areal data under different sampling scenarios and with two different cases for areal and point sample sizes.

The second part studies how much prediction error is incurred when using the alternative hierarchical model instead of the original formulation. Recall that this model is preferred for computation purposes, even though it is known that areal data is not being handled appropriately. We know that a very fine triangulation of the region will produce better results, but at a higher computational cost. To be able to use a more sparse mesh we need to forgo the requirement that each areal block must include at least one vertex, and that's why we use the hierarchical model formulation instead.

1.4.1 Sampling Scenarios

The main objective of this simulation study is to assess how well INLA is able to combine both point and areal data under different data sampling scenarios and with different sample sizes. For this, 100 'true' fields are simulated from mean-zero Gaussian process with Matérn covariance function.

The covariance function uses $\nu = 1, \sigma^2 = 1$ and $\rho = 142$. For simplicity in the simulation, the mean function is taken to be fixed at $\mu = 0$. Each field is generated on a 100×200 regular grid, producing 20,000 true points.

This model will eventually be fitted for measuring global temperatures, which are values vastly different from what we're simulating, but for this simulation study we are more interested in seeing the effects of how the data is sampled, as this is the most important characteristic of the real data.

We will study five different sampling scenarios under two different ways of splitting the point and areal sample sizes; the sampling scenarios are summarized in Table 1.1. Point data is sampled at random from the original field in one of two cases: either fully at random over the entire domain or constrained inside land masses. When sampled inside land masses, they are further sampled such that the northern latitudes have more points than the southern ones. This replicates the pattern observed in the real data where the point data is mostly in land and with more prevalence in Northern America and Europe. In either case, point data is also sampled with an added measurement error as seen in Equation (1.1).

Scenario	Areal Data	Point Data
1	All data	Sampled at random
2	All data	Sampled in land
3	Missing data at random	Sampled at random
4	Missing data in stripes	Sampled at random
5	Missing data in stripes	Sampled in land

Table 1.1: Different sampling scenarios for the areal and point data. Areal data can either be fully present, or some cells might be missing at random or in stripped patterns. Point data can either be sampled at random over the entire domain, or constrained to only land masses, with more prevalence in northern latitudes.

In the case of areal data, we generate a 90×180 grid, where each block averages 1,2 or 4 points from the original field, following the form of (1.2). When 'sampling' areal data we're actually considering how we're sampling blocks to be considered as missing data. Scenarios 1 and 2 use the whole grid of areal data. Scenario 3 samples the missing blocks at random over the entire spatial domain. Scenarios 4 and 5 sample the missing blocks inside 'stripes'. These stripes are generated using a sinusoidal function on longitude and latitude that is tuned to produce diagonal vertical stripes (as seen in Figure 1.3). The blocks that are removed correspond to the highest values of the sine wave, according to another user-tuned threshold. All the scenarios that have missing data have

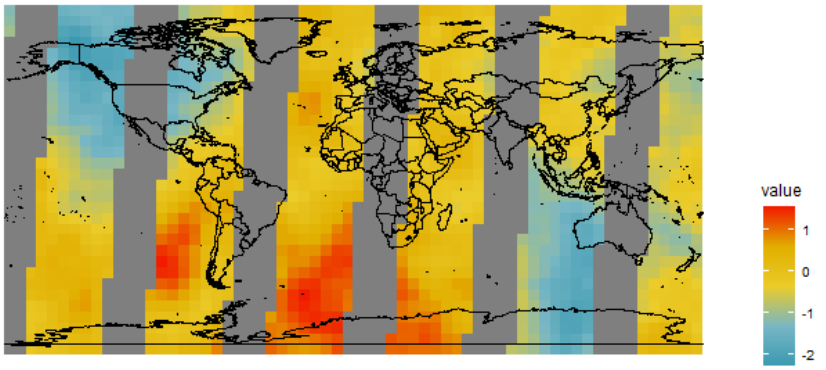


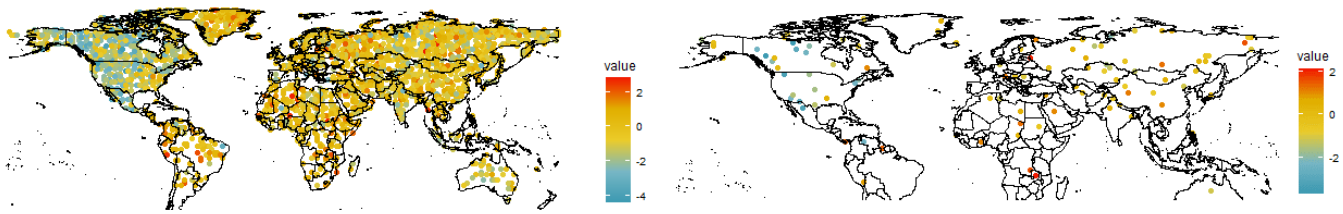
Figure 1.3: Example of a simulated areal dataset with missing blocks in a vertically diagonal stripe pattern.

the same amount of removed data.

The sample sizes are considered in two different splits of how much point data we want to have in comparison to areal data. Areal data is always kept fixed at 16,200 blocks (without missing data). As for point data, one case considers having the same number of point data as areal data, whereas the other case is more similar to the real data and considers that point data represents about 5% of the total data; these splits are referred to as the '50-50' and '95-05' splits, respectively. Figure 1.4 shows an example of point data sampled in land under both sample size splits.

The sampling scenario that resembles the most the real data is Scenario 5 under the 95-05 split. All other Scenarios are considered as they represent 'better' versions of what we observe in the real data and we wish to compare how the predictions under Scenario 5 compare to the rest. Similarly, the 50-50 split also represents a more optimistic situation where we have equal amounts of data from both sources.

The model that properly treats areal data as such was fitted for each combination of sampling scenario and sample size split. The metrics observed for prediction accuracy were the root mean square error and the correlation between predicted values and true values at the original 100×200 resolution. The predicted values consist of the posterior means of the fitted value. RMSE results are shown in Figure 1.6, whereas the correlation results are presented in Figure 1.6.



(a) (b)

Figure 1.4: Point data sampled in land, with more data in northern latitudes, under the (a) 50-50 split and the (b) 95-05 split of areal to point data sample sizes.

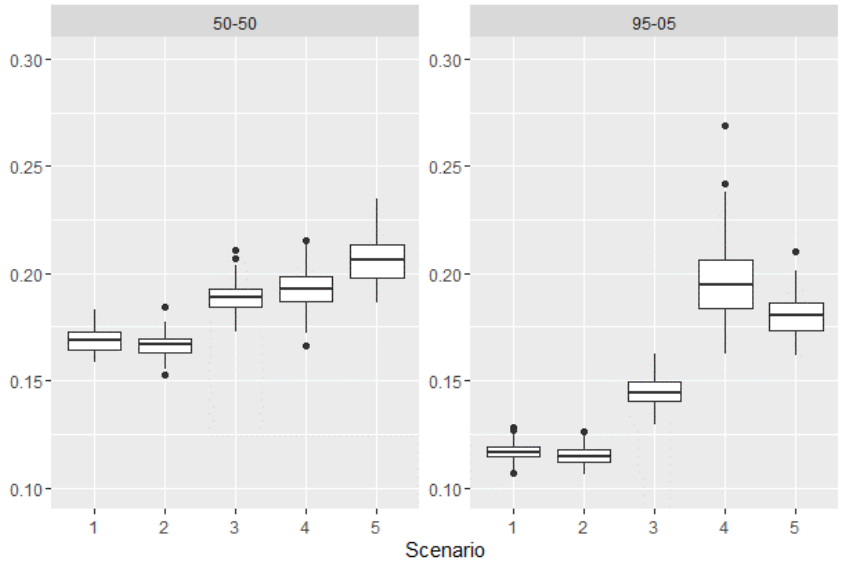


Figure 1.5: Root mean square error between the true and predicted values of $S(\boldsymbol{x})$ for each of the five different sampling scenarios and the two sample size splits.

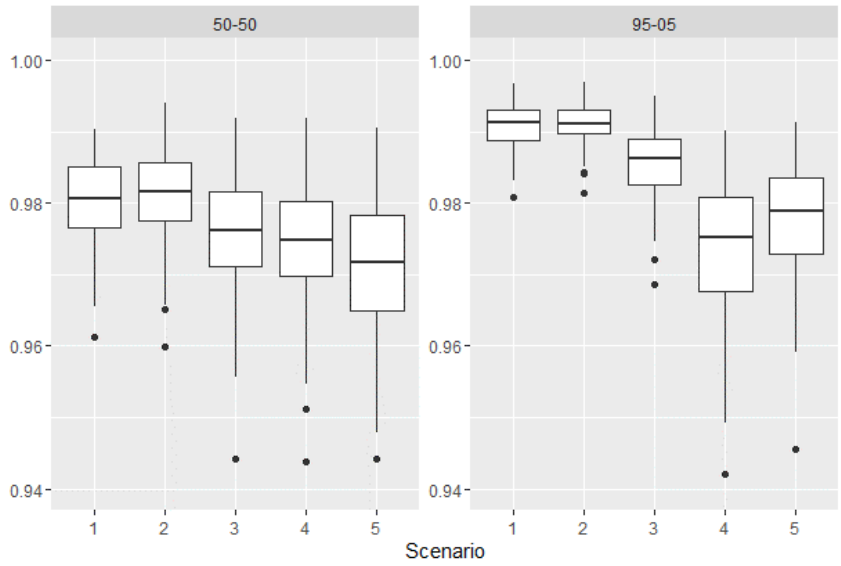
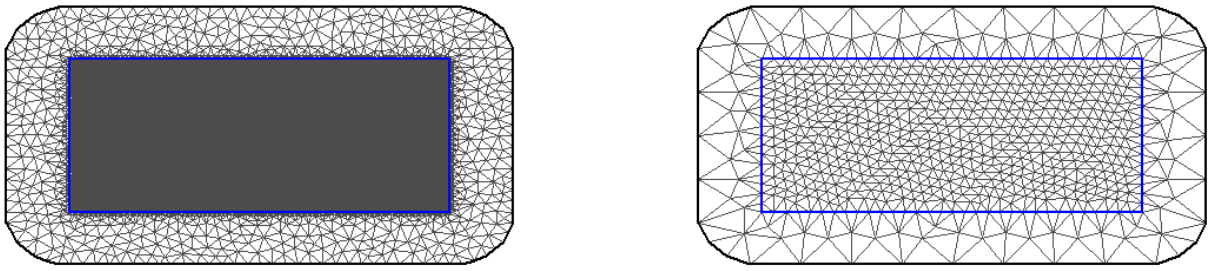


Figure 1.6: Correlation between the true and predicted values of $S(\boldsymbol{x})$ for each of the five different sampling scenarios and the two sample size splits.

From both plots we can observe that high correlations relate to low RMSEs and vice-versa. In general, we observe that results in the 95-05 split are better than in the 50-50 split, which at first seems counter-intuitive as one would expect more data to be better for the analysis, but we must consider that the only source of noise in these simulations come from point data. Further simulations with reduced error in the point data (not shown) indeed show under the presence of missing areal data, predictions are better for the 50-50 split than the 95-05 split. Furthermore, under the presence of all the areal data, both splits coincide that having point data throughout the entire region is better than clustered in land.

When introducing missing blocks in areal data, it is clear that having them be missing at random (i.e. Scenario 3) is better than having them missing in the striped pattern (Scenarios 4 and 5). As for the sampling of point data with striped areal missing data, the results show an interesting reversal between the different splits. Under the 50-50 split, having points sampled in the whole domain is better than just sampling them in land, but the reverse is true for the 95-05 split. This once again seems to be due to noise being present only in point data, since when this noise is removed (not shown), the trend of the 50-50 split is also observed in the 95-05 split.

Overall, the simulation study seems that suggest that under the presence of noisy point data (or at least noisier than areal data), having less point data is better, irregardless of whether these



(a) (b)

Figure 1.7: Delaunay triangulation of the study region (inside the blue polygon) using a (a) dense mesh and a (b) sparse mesh. The dense mesh has 18066 vertices, whereas the sparse mesh consists of only 708, much less than the total sample size (16302).

points are sampled in the whole domain or just in land. Nevertheless, it is clear from both splits that the best case scenario would be having a full areal data set, but in practice this is not feasible.

1.4.2 Using the hierarchical model

This study aims to quantify the loss in prediction accuracy when using the hierarchical model (1.4), as opposed to the original formulation in (1.1) and (1.2). Recall that the hierarchical model is known to not be correct, but can be fitted using a much sparser mesh and thus will be much faster to compute. Other simulations done (not shown) revealed that as the number of areal blocks increased, it didn't matter if they were treated as areal or point data, as long as the mesh was kept the same. This study hopes to asses how a sparser mesh would affect the predictions, with using the alternative method as a counterbalance to just treating areal observations as simple point data.

For this simulation we use only one of the simulated fields and sample point and areal data in a manner equivalent to Scenario 2 under the 95-05 sample size split. We then proceed to fit the model using both formulations, with the second formulation using a much sparser mesh. Both meshes can be seen in Figure 1.7, where it is very difficult to observe the small triangles inside the study domain (delineated by the blue line) for the dense mesh.

The estimated fields are shown in Figure 1.8. Visually, there is almost no difference between

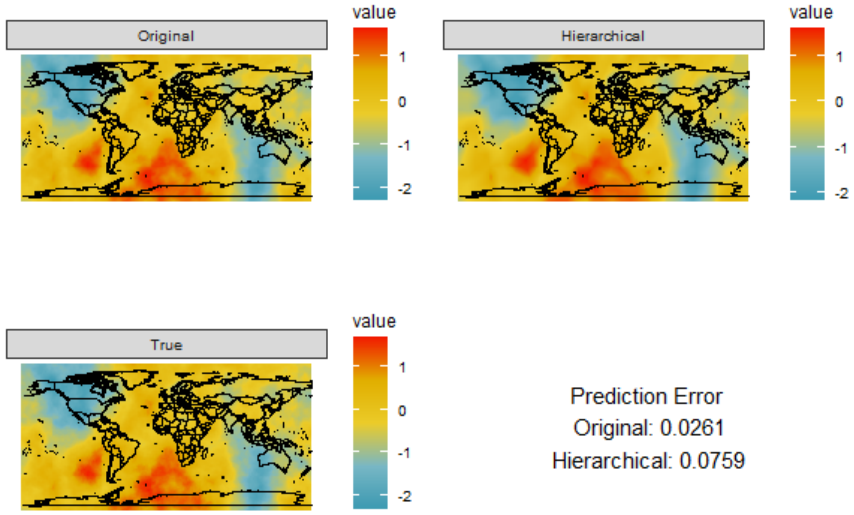


Figure 1.8: Estimated fields for the original model formulation that accurately treats areal data and for the hierarchical model formulation that treats areal as point data with a different error structure. The true field and the prediction errors (RMSE) are also included for reference.

both estimated fields and the truth, except for some smoothing in the results. However, the root mean square error between the true values and the estimated ones are almost three times as high when treating areal data as point data in the alternative model, than when treating it appropriately in the original formulation. Nonetheless, the error is still relatively low.

In terms of correlation, the correlation between the estimation and the true field for the original formulation was 0.999, whereas the correlation with the alternative model was 0.995, which is not a terrible loss. Nonetheless, the computation time for the original method was around 35 minutes, whereas the alternative model took about 8 seconds, which is a drastic difference.

Figure 1.9 shows the posterior standard deviations for the fitted values under the two different modeling approaches. As seen in the hierarchical model plot, there are spots scattered around the domain that correspond to higher standard deviations. These lie close to the vertices of the sparse mesh. Overall, the mean standard deviation for the original formulation is about 0.002, whereas for the alternative model it is 0.014.

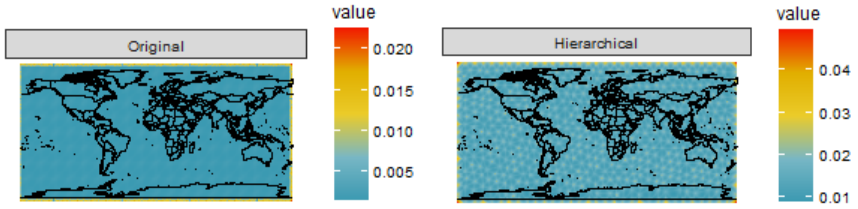


Figure 1.9: Posterior standard deviations for each fitted value for the original model formulation as well as the alternative hierarchical model.

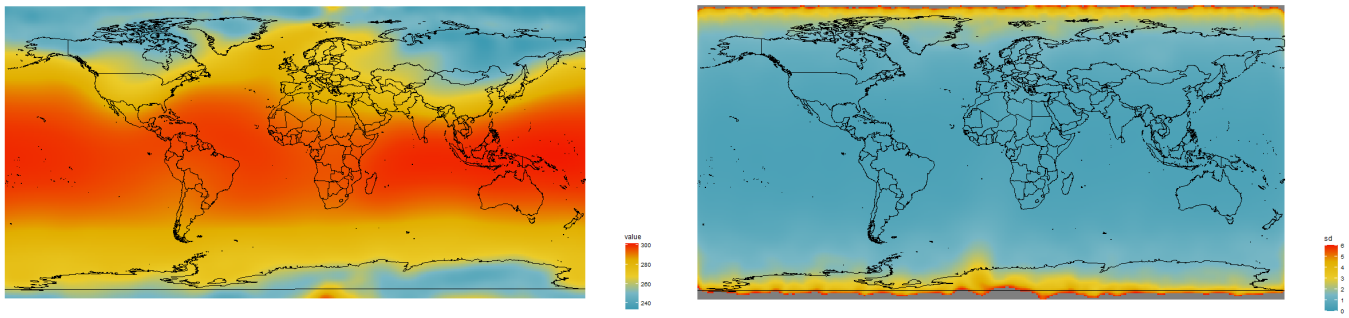
1.5 Interpolation of global temperature

We will only present the results pertaining to January 1st, 1990 here. The data is fitted using both the original formulation in equations (1.1) and (1.2), as well as the modified hierarchical model (1.4). The original formulation uses a dense mesh, similar to the one in Figure 1.7a, whereas the alternative model uses the same sparse mesh in Figure 1.7b.

Results for the model that uses the dense mesh are presented in Figure 1.10. The estimates of the field are the posterior means of the fitted values at the centroids of each areal block and they are given in Kelvin. This model took approximately 45 minutes to run. Figure 1.11 shows the results for the alternative model with the sparse mesh. These results were obtained in about 30 seconds.

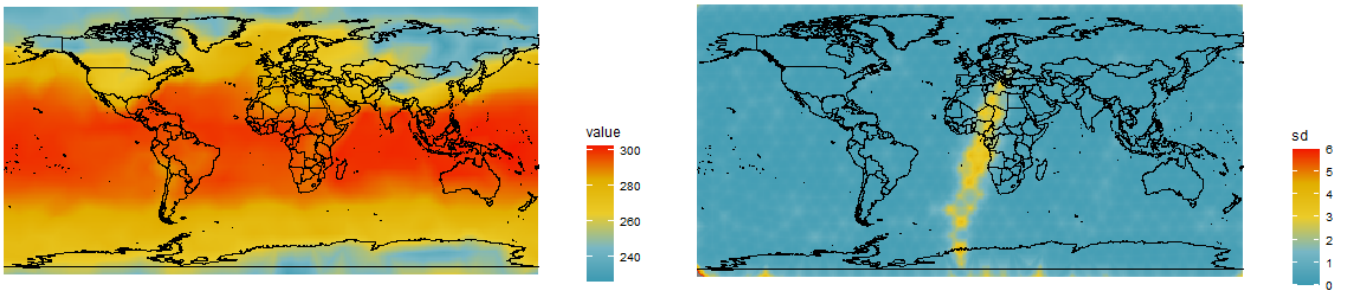
Both estimated fields look very similar in general terms. They seem to do well appropriately capturing the regions of cold and hot climate for this particular time of the year. Interestingly, the original model seems to estimate a small region of hot temperatures in Antarctica that is unexpected.

The difference in meshes is more noticeable when looking at the posterior standard deviations. Standard deviations in Figure 1.10b are overall smaller than those in 1.11b. Nevertheless, there is higher uncertainty in regions where areal data was missing (although, due to the presence of high values in the border, this is difficult to see in the results for the original model). Figure 1.11b also



(a) (b)

Figure 1.10: Results obtained by treating point and areal data accordingly. The field estimate (a) correspond to the posterior means and the standard deviation (b) to the posterior standard deviations at each location.



(a) (b)

Figure 1.11: Results obtained by treating point and areal data as point data but with different error structures. The field estimate (a) correspond to the posterior means and the standard deviation (b) to the posterior standard deviations at each location.

shows small regions of high uncertainty scattered around the map that correspond to the vertices of the sparse mesh.

Both results show higher uncertainty around the border. This is a common problem with the SPDE approach that is mitigated by having larger triangles outside the study region. However, this is still difficult to mitigate when the inner triangles are very small, as the outer triangles grow in size starting from the size of the inner ones. This means that the triangles that are closest to the mesh are still very small (as seen in Figure 1.7a) and thus produce higher border uncertainties.

1.6 Conclusion

INLA is capable of accurately combining areal and point data through the SPDE approach. However, this is heavily dependent on the mesh, as areal data requires at least one mesh vertex inside of each areal block. This becomes a issue when areal blocks become smaller as computations take much longer. We propose an alternative hierarchical model that relates areal and point data to the same latent process but allows flexibility in modeling the error of each type of data. This allows the GMRF to be fit in a more sparse network of vertices and drastically reduce computation times.

This project presents an interesting question moving forward. Are we more interested in the higher accuracy by treating areal and point data accordingly, at the cost of longer computation times, or can we sacrifice some of that accuracy to have the results out faster? Simulations show that there is a loss in accuracy by using the faster method, but this loss is decreased as the number of areal blocks become larger.

Through a simulation study that changes the way areal and point data are obtained we have seen that our results would be benefited from having less missing areas coming from TOVS. They also seem to suggest that point data is not as essential as areal data, but this is only true if point data is particularly noisy. As radiosonde and weather balloons become better measuring instruments we can assume that measurement error would be decreased and their presence would help with the final predictions much more.

There is one major problem when using this method to interpolate world data, being that the mesh is constructed in a 2D surface, as opposed to a sphere. Lindgren, Rue, and Lindström 2011 shows that it is possible to create a mesh in a sphere, but as far as we know this is still not

implemented into R-INLA. This means that the estimated covariance structure is dependent on the projection of the world we use. This is most likely the reason why regions near each of the poles seem to have higher temperatures than expected, as the covariance structure identifies these regions to be farther away than what they should be.

As previously mentioned, this is part of a bigger project that is trying to find the best way to handle the change of support problem in terms of both accuracy and computation time when it comes to integrating observed datasets. Additionally, these methods should also be able to provide some quantification of prediction error, which in this case can be handled by the posterior standard deviations. So far, the results for world temperature are promising and sensible. In the future, we would like to apply a similar methodology to aerosol data, which pose an extra set of challenges with how that data was recovered in the early 90s.

References

Anyamba, Ebby and Joel Susskind (1998). "A comparison of TOVS ocean skin and surface air temperatures with other data sets". In: *Journal of Geophysical Research: Oceans* 103.C5, pp. 10489–10511.

Durre, Imke et al. (2016). "Integrated Global Radiosonde Archive (IGRA), Version 2". In: *NOAA National Centers for Environmental Information*. DOI: 10.7289/V5X63K0Q.

Lindgren, Finn and Håvard Rue (2015). "Bayesian Spatial Modelling with R-INLA". In: *Journal of Statistical Software* 63.19. DOI: 10.18637/jss.v063.i19.

Lindgren, Finn, Håvard Rue, and Johan Lindström (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.4, 423–498. DOI: 10.1111/j.1467-9868.2011.00777.x.

Moraga, Paula et al. (2017). "A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE". In: *Spatial Statistics* 21, pp. 27–41.

Rue, Håvard, Sara Martino, and Nicolas Chopin (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.2, 319–392. DOI: 10.1111/j.1467-9868.2008.00700.x.