Temperature prediction analysis considering an optimal distribution of stations: clusters or randomly

Pablo Juan[1], Laura Serra[2,4,*], Diego Varga[2,3] and Marc Saez[2,4]

[1] *Department of Mathematics, University Jaume I, Castellón, Spain; juan@mat.uji.es*
[2] *Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, Spain; laura.serra@udg.edu, dievarga@gmail.com, marc.saez@udg.edu*
[3] *Geographic Information Technologies and Environmental Research Group, University of Girona, Spain*
[4] *CIBER of Epidemiology and Public Health (CIBERESP), University of Girona, Spain*
[*] *Corresponding author*

Abstract.
Global mean surface air temperature is the most used measure of the climate system. Nowadays, due to the climate change problem, the interest of predicting climatic values in areas without stations has increased a lot and has been developed new interpolation methods. If we associate temperature stations with their spatial coordinates, along with other variables, it is possible to identify them by means of a spatio-temporal stochastic process. Two are the objectives in this work. Firstly, to predict the mean temperature throughout Catalonia taking into account the total number of stations (180)*. Secondly, to analyse the goodness of prediction reducing the number of stations gradually. At first, we consider less randomly chosen stations* $(160, 100, 80)$ *and then we select stations, which are in clusters. We specified spatial log-Gaussian process models. Models are estimated using Bayesian inference for Gaussian Markov Random Field (GMRF) through the Integrated Nested Laplace Approximation (INLA) algorithm. The results allow us to quantify the minimum number of stations which are needed to do the best prediction of the mean temperature in Catalonia as well as to know the best distribution of these stations. We believe the methods shown in this study may contribute to improve prediction studies and to reduce computational cost of these predictions.*

Keywords. *Covariates, Prediction, SPDE, GMRF, INLA.*

---

1 Introduction

Temperature is one of the most important atmospheric variables which directly impact physical and biological processes. Climate is only one of the many physical variables which impact life on Earth. However, one of the major concerns with a potential change in climate is that an increase in extreme

events will occur. For this reason, interpolation methods of climatic data have been widely studied. Specifically, in recent years, due to the growing interest that IPCC reports and the Kyoto Protocol have caused new methodologies have been developing. The spatial availability of climate data can be a problem because, although the information is recorded, weather station network is often sparse. In this way, different interpolation methods have been developed to predict climatic values in areas without stations. If we associate temperature stations with their spatial coordinates, along with other variables, it is possible to identify them by means of a spatio stochastic process. In fact, what is usually of interest is to assess their dependence on covariates. In order to optimize the number of stations used to get the best prediction of the mean temperature it is necessary to distinguish between two sources of extra variability; the largest source usually named 'spatial dependence' or clustering and the one called uncorrelated or non-spatial heterogeneity, which is due to unobserved non-spatial variables that could influence the dependent variable (Lawson et al, 2003; Barceló et al, 2009). To take into account the spatio variability, we introduce some structure into the model. In particular, we follow the recent work of Lindgren (Lindgren et al. 2011), and specify a Matérn structure (Simpson et al. 2011). In short, we use a representation of the Gaussian Markov Random Field (GMRF) explicitly constructed through stochastic partial differential equations (SPDE) which has as a solution a Gaussian Field (GF) with a Matérn covariance function. To sum up, instead of using the Matérn in a regular lattice, which is the usual practice and would imply an estimation with a high computational cost as well as one that would be weak in terms of efficiency (Lindgren et al. 2011), we specify the structure of the spatial Matérn covariance in a triangulation (Delaunay triangulation) of the studied area with a low computational cost and, more importantly in our context, much greater efficiency. Therefore, in this work we apply a computationally efficient approach based on the stochastic partial differential equation (SPDE) models and we use SPDE to transform the initial Gaussian Field (GF) to a Gaussian Markov Random Random Field (GMRF). GMRFs are defined by sparse matrices that allow for computationally effective numerically methods. Furthermore using Bayesian inference for GMRFs, it is possible to adopt the INLA algorithm that gives significant computational advantages.

2 Method

2.1 Data setting

In this study we consider 180 stations homogeneously distributed throughout the territory. In addition to the locations of the stations centroids, measured in Cartesian coordinates (Mercator transversal projections, UTM, Datum ETRS89, zone 31-N), some spatial covariates are considered. In particular, we use the relative humidity and elevation.

2.2 Statistical models

In statistical analysis, to estimate a general model is useful to model the mean for the i-th unit by means of an additive linear predictor, defined on a suitable scale

$$\eta_i = \alpha + \sum_{m=1}^{M} \beta_m z_{mi} + \sum_{l=1}^{L} f_l(\nu_{li}) \tag{1}$$

In our case, assuming that the subscript i denotes the mean temperature in a particular geographical area, we specify the log-intensity of the Poisson processes by a linear predictor (Illian et al., 2012) of the form

$$\eta_i(s_j) = \beta_0 + \beta_1 X_j + \beta_2 Z_j + S_j \tag{2}$$

where S_j is the spatial dependence. Given the specification in (2), the vector of parameters is represented by $\theta_j = \{\beta_0, S\}$ where we can consider $X_i = (S)$ as the i-th realization of the latent GF $X(s)$ with the Matérn spatial covariance function. We can assume a GMRF prior on θ , with mean 0 and a precision matrix Q. In addition, because of the conditional independence relationship implied by the GMRF, the vector of the hyper-parameters $\psi = (\psi_s)$ will typically have a dimension of order much smaller than θ. Models are estimated using Bayesian inference for Gaussian Markov Random Field (GMRF) through the Integrated Nested Laplace Approximation (INLA). The use of INLA and the SPDE algorithms produce massive savings in computational times and allow the user to work with relatively complex models in an efficient way. Once the model is estimated, we predict the mean temperature considering different number of stations and then we use stations spatial distributed according different criteria. Prediction's results are tested analysing the minimum Deviance Information Criterion (DIC) and the lowest mean error. A prediction map is created with the best results. All analyses are carried out using the R freeware statistical package (version 2.14.1) (R-Development Core Team, 2011) and the R-INLA package (R-INLA, 2012).

3 Results

Here we represent how stations are distributed through the studied area.

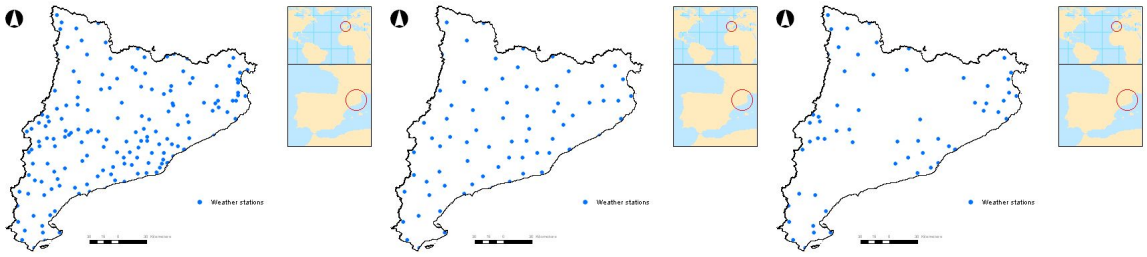

Figure 1: From left to right: all the stations; half of the stations and the cluster effect.

Considering all the stations we construct prediction maps of the mean temperature such as the following:

Acknowledgments. We would like to thank the Environment Department of the Government of Catalonia for the access to the digital map databases.

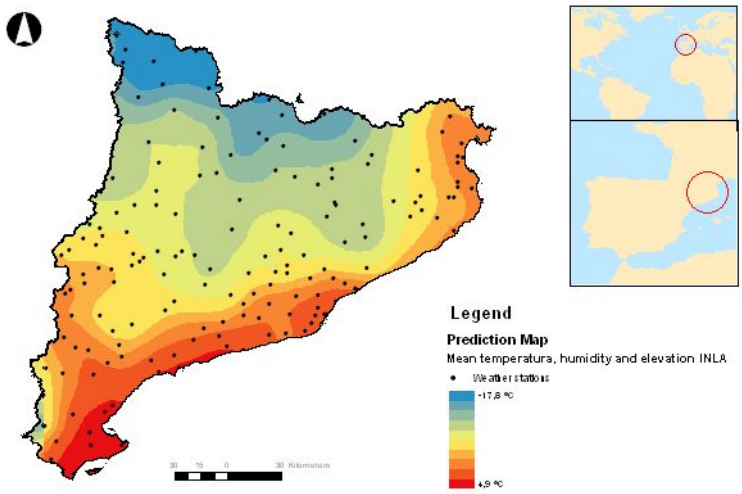

Figure 2: Predicton map of mean temperature.

References

[1] Illian, J.B., Sørbye, S.H., Rue, H., Hendrichsen, D. (2010) Fitting a log Gaussian Cox process with temporally varying effects - a case study, http://www.math.ntnu.no/inla/r-inla.org/papers/S17-2010.pdf.

[2] Lindgren, F., Rue, H., Lindstrom, J. An explicit link between Gaussian fields and Gaussian Markov random fields: the SPDE approach (with discussion). *Journal of the Royal Statistical Society, Series B*, **73 (4)**, pp.423-498, 2011.

[3] R-INLA project (2011). URL: http://www.r-inla.org/home.

[4] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, 2011, URL http://www.R-project.org/

[5] Saez, M., Barceló, M.A., Tobias, A., Varga, D., Ocaña-Riola, R., Juan, P., Mateu. J. (2012).Space-time interpolation of daily air temperatures. *Journal of Environmental Statistics* **3 (5)**.

[6] Simpson, D., Illian, J., Lindgren, F., Sørbye, S.H., Rue, H. Going off grid: Computationally efficient inference for log-Gaussian Cox processes, 2011, http://www.math.ntnu.no/ daniesi/S10-2011.pdf