Evaluating Spatio-temporal models for crop yield forecasting using INLA: implications to pricing area yield crop insurance contracts

Ramiro Ruiz-Cárdenas[1]* and Elias Teixeira Krainski[2]

[1] Laboratório de Estatística Espacial, Universidade Federal de Minas Gerais - Brazil
[2] Departmento de Estatística, Universidade Federal do Paraná - Brazil

1 Introduction

Area yield crop insurance is a recent insurance product, in which farmers collect an indemnity whenever the county average yield falls beneath a yield guarantee, regardless of the farmers actual yields. The pricing methodology for this kind of insurance requires the estimation of the expected crop yield at the county level. This can be done in a hierarchical Bayesian framework via spatio-temporal modelling of areal crop yield data, which allows estimates of the premium rates be obtained directly from the posterior predictive distribution of crop yields, capturing inference uncertainties involved in predicting the insurance premium rates. Inference in this kind of models is typically based on Markov chain Monte Carlo methods (MCMC), a computer-intensive simulation-based approach. However, these methods suffer from several problems: Computational time is long, parameter samples can be highly correlated and estimates may have a large Monte Carlo error. Additionally, a huge number of models with different components resulting from the combination of regional effects, time trends and time-space interactions, as well as of several covariates entering in the models in different ways, need to be fitted and compared in order to identify the more suitable one to be used in the calculation of the premium rates of the areal crop yield insurance contract. This task becomes very time consuming when the number of areas increases.

A recent alternative to MCMC methods to perform inference in latent Gaussian models are the integrated nested Laplace approximations (INLA) (Rue et al., 2009). The methodology is best suited for latent Gaussian models specified as Gaussian Markov random fields (GMRF). In this paper the INLA approach is applied to fit and compare in an efficient way several spatio-temporal hierarchical crop yield models in order to identify the most suitable ones to calculate the premium rates of an areal crop yield insurance contract for maize in Paraná state (Brazil). Further we propose an extension of the INLA approach that enable the application of the same methodology using a complex dynamic spatio-temporal model for areal data within reasonable computational time and in a user friendly way with INLA.

The remainder of the paper is organized as follows: In Section 2 we outline the INLA computational approach. Section 3 describes in detail the methodology proposed to fit and compare several spatio-temporal models using INLA under a Bayesian hierarchical framework. In Section 4 we present the main results for the models considered in section 3. Current

*Corresponding author. Email: ramiro_rc1@yahoo.com.br

extensions and future work are discussed in Section 5, where we propose an extension of the INLA approach to perform approximate inference in a complex dynamic spatio-temporal model for crop yield forecasting.

2 Integrated Nested Laplace Approximations (INLA)

INLA is a computational approach to perform approximate Bayesian inference based on an efficient combination of Laplace approximations and numerical integration. Unlike MCMC, the INLA method does not sample from the posterior distribution. It approximates the posterior with a closed form expression. Therefore, problems of convergence and mixing are not an issue. According to Schrödle and Held (2009), the method is best suited to Bayesian hierarchical models for which there is a large number of unknown parameters following a Gaussian Markov random field denoted as $\pi(\boldsymbol{x} \mid \boldsymbol{\theta})$ and a small number of hyperparameters, with a specific form of prior covariance on the parameters. In the following, the way how INLA computes posterior marginal distributions of parameters of interest is described in brief. For details see Rue et al. (2009).

Following the notation of Schrödle and Held (2009), let $\boldsymbol{x}$ denote the vector of all Gaussian variables and $\boldsymbol{\theta}$ the vector of hyperparameters, which are not necessarily Gaussian. The main goal of a Bayesian inference method in the proposed setting is to estimate the posterior distribution

$$\pi(x_i \mid \boldsymbol{Y}) = \int_{\boldsymbol{\theta}} \pi(x_i \mid \boldsymbol{\theta}, \boldsymbol{Y}) \pi(\boldsymbol{\theta} \mid \boldsymbol{Y}) d\boldsymbol{\theta} \tag{1}$$

given the data for each component x_i of the Gaussian field $\boldsymbol{x}$. The number of components of $\boldsymbol{\theta}$ should not be too large for accurate inference (since these components are integrated out via Cartesian product numerical integration, which does not scale well with dimension). The key feature of the INLA approach is to construct a nested approximation for (1).

The second component in the integral (1), the marginal posterior density $\pi(\boldsymbol{\theta} \mid \boldsymbol{Y})$ of the hyperparameters $\boldsymbol{\theta}$, can be approximated by

$$\tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y}) \propto \left. \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{Y})}{\tilde{\pi}_G(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{Y})} \right|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})} \tag{2}$$

where $\tilde{\pi}_G(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{Y})$ denotes the Gaussian approximation to the full conditional distribution of $\boldsymbol{x}$ (Rue and Held, 2005) and $\boldsymbol{x}^*(\boldsymbol{\theta})$ is the mode of the full conditional of $\boldsymbol{x}$ for a given $\boldsymbol{\theta}$. The main use of $\tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y})$ is to numerically integrate out the uncertainty with respect to $\boldsymbol{\theta}$ in (1). It is important to find good support points θ_k, $k \in \{1, ..., K\}$, for a numerical integration of (1). To produce these grid of points, the mode of $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})$ is located, and the Hessian is approximated, from which the grid is created and exploited in (1).

Three approaches were proposed by Rue et al. (2009) to approximate the first component $\pi(x_i \mid \boldsymbol{\theta}, \boldsymbol{Y})$ of the integral in (1): A Gaussian approximation, a full Laplace approximation and a simplified Laplace approximation. Each approach has different features and the results are supposed to be differently accurate. The simplest approximation is the Gaussian approximation, which gives quite satisfactory results in short computational time. An advantage of the Gaussian approximation is that it is also straightforward to correct for linear constraints imposed on the latent field $\boldsymbol{x}$. However, there can be numerical errors in the location and/or errors due to the lack of skewness of the Gaussian approximation. It can be improved through applying another Laplace approximation to $\pi(x_i \mid \boldsymbol{Y})$. This "full Laplace" approximation is supposed to be most accurate. An alternative called "simplified Laplace" approximation, is less expensive from a computational point of view with only a slight loss of accuracy. This method works fine in terms of approximation error for many observational models and is based on a series expansion of the full Laplace approximation. However, it is not so straightforward to incorporate linear constraints on $\boldsymbol{x}$ in the full Laplace approximation and its simplified version. This is due to the fact that both approximations work directly with the posterior marginals of the components x_i of $\boldsymbol{x}$ in turn, not taking into account deterministic dependencies between components of $\boldsymbol{x}$. Linear constraints on $\boldsymbol{x}$ are therefore not fully incorporated in the improved estimates of the posterior marginals. Thus, $\pi(x_i \mid \boldsymbol{Y})$ may be evaluated via the approximation

$$\tilde{\pi}(x_i|\boldsymbol{Y}) \approx \sum_{k=1}^{K} \tilde{\pi}(x_i|\theta_k, \boldsymbol{Y}) \times \tilde{\pi}(\theta_k|\boldsymbol{Y}) \times \Delta_k.$$

For substitution of the integral in (1) an area weight Δ_k has to be assigned to each θ_k. Its size depends on the actual strategy of choosing the θ_k's. The output of INLA consists of posterior marginal distributions, which can be summarized via means, variances and quantiles.

Approximate inference in this work was performed with the R programming language (R Development Core Team, 2009) using the INLA R-package available at `http://www.r-inla.org`.

3 Methodology

The choice of a statistical model that adequately reflects the conditional density of crop yields is an important consideration in the actuarial calculation of an accurate premium rate. In doing this, a number of issues relating to the modeling of crop yields must be considered, such as the fact that crop yield series tend to have substantial trends and tend to be significantly correlated across space due to the systemic nature of weather.

The terms of a crop insurance contract typically must be available one to two years prior to the insurance cycle. This reflects the fact that information on areal crop yield may take some time to be adequately measured and corrected before it can be released (Ozaki et al., 2008). Further, an insurer will not offer coverage after the insurance buyers have information about their yields. For example, contracts must typically be signed before sowing season. Otherwise, farmers would have an information advantage over insurers, who had to specify contract parameters at a much earlier date. In addition, administrative issues related to the operation of insurance programs usually lead to a substantial delay in the specification of contract parameters. We will assume here that there is a two-year lag between the receipt of historical yield data and the deadline required for filling new contract terms. In this context, we must attempt to choose the best possible statistical model to predict yields for the following two years.

The available data set consisted of average annual county yield records for corn in 397 counties of Paraná state during the period 1980–2008. This data set was provided by the Brazilian Institute of Statistics and Geography (IBGE) and by the Agriculture office of Paraná State (SEAB). Daily weather data (rainfall, maximum and minimum temperature and solar radiation), needed to construct the covariates, were also available for the same period at 87 weather stations spread across the state. All the space-time models were implemented using data for the period 1980–2006 ($T = 26$), leaving out the last two years (2007–2008) to compare the actual values with those predicted by the models one and two-steps ahead (that is, at $T + 1$ and $T + 2$).

In this analysis we consider a parametric modeling approach, and assume that crop yields tend to follow a normal distribution (Just and Eininger, 1999). We adopt a hierarchical Bayesian inferential framework that accounts for all sources of uncertainty. Approximate inference for all models considered is performed using the INLA approach. Following Ozaki et al. (2008), and as usually in crop insurance contracts, we assume that "best production practices" are followed (i.e., that no moral hazard exists), and thus optimal levels of input usage are assumed and yields are not typically conditioned on inputs. Next subsection describes the Bayesian hierarchical space-time models considered and the comparison procedure among them. A dynamic spatio-temporal modelling approach to crop yield forecasting is outlined in section 5.

3.1 Hierarchical space-time crop yield models

Let y_{it} be the corn yield registered at time t, $t \in \{1, \cdots, T\}$ in county i, $i \in \{1, \cdots, S\}$, where $S = 397$ and $T = 26$. A general version of the hierarchical space-time models considered in this analysis can be represented by the following hierarchical structure, where we simultaneously model the time trend and the temporal and spatial autocorrelation:

$$\begin{aligned} y_{it} &\sim N(\mu_{it}, \tau_j), \qquad , \; j \in \{1, \cdots, 5\} \\ \mu_{it} &= \rho_i y_{i,t-1} + \beta_{0_i} + \beta_{1_i} t + \beta_{2_i} t^2 + \sum_{z=3}^{Z} \beta_{z_i} \, \xi_{zit}. \end{aligned}$$

Here $\rho_i \sim N(\alpha_\rho, \tau_\rho)$, $\alpha_\rho \sim N(0, \tau_\alpha)$, τ_ρ and τ_j, $j \in \{1, \cdots, 5\}$ are precision parameters following a priori an inverse gamma distribution and ξ_{zit} represents the z-th covariate for area i at time t. In this model it is assumed that the variance τ_j is different in each of five "macroregions" previously defined.

Several sub-models can be fitted from this general model, considering different ways of entering with each of the terms and covariates in the model. The schematic representation of all possible ways in which these terms could enter in the models is shown in Figures 1 and 2. As possible covariates we consider the planted area in each county (in Hectares) and the aggregated crop yield at the "nucleo regional" level (the Paraná state divided in 20 areas). Information on crop yields for the forecasting period (T+1 e T+2) is available for this level of aggregation from the Agricultural Office of Paraná state. Three agroclimatic indices were also considered as covariates: the Standardized Actual Evapotranspiration Index (IPER), the Water Requirement Satisfaction Index (WRSI) and the Standardized Precipitation Index (SPI). The WRSI is based on the ratio of actual and potential evapotranspiration (ETa/ETp)[1] accumulated during the critical period of the crop in terms of water deficit. These quantities are determined by the daily soil water balance from appropriate meteorological and soil input data. The SPI (McKee et al., 1993) is a drought index based on the transformation of the rainfall depth for a given window size to its corresponding cumulative probability, which is then mapped onto the standard normal scale. Therefore, the actual rainfall can be expressed as a standardized departure from the rainfall probability distribution function. The IPER (Blain and Brunini, 2006), is an adaptation of the SPI methodology for quantifying agricultural drought in a 10-days scale, based on the fit of the actual evapotranspiration series to the beta distribution (for details see Blain and Brunini, 2006).

Farmers in a same county may sow different maize varieties in different dates. Therefore, the determination of a unique critical period for maize (in terms of water requirements), which can be considered representative of all farmers in each county, to calculate the agroclimatic indices above, is difficult to establish. Instead, each of these indices were accumulated through

[1] *ETa* is the amount of water lost by the crop as a function of climate, the plant processes and the availability of water in the soil. It is calculated from the daily water balance. *ETp* is the maximum consumption of water by a crop under optimal conditions and full availability of water in soil.

six subperiods of 20 days each, covering all the phenological phases of maize and all possible combinations of these subperiods were evaluated as covariates in the models. This is important, as sowing dates follow a spatial pattern in the sense that in some regions of the state farmers begin sowing earlier than in others. Therefore, a given critical subperiod may be not important for some regions but it may be for others.

In a first stage we evaluated models with and without the autoregressive term and with all combinations of temporal trends as represented in Figure 1, and keeping the covariates either absent or fixed and joint (that is, the six subperiods at a time) in the models. The best models obtained in this first approach were considered for a second stage evaluation, where models with all possible combinations of covariates, as described in Figure 2, were fitted and compared.

The time trend coefficients (β_0, β_1 and β_2) could be either, fixed for all the areas or varying across the areas. This variation could be at the county level (397 areas) or at a the more aggregated "microregion" level (39 areas). These coefficients also could have a spatial structure or to vary in a random way. It was assumed that the variation of each area for spatially structured coefficients follows a priori an intrinsic CAR distribution (hereinafter denoted as BESAG), that is:

$$\beta_{j_i} \mid \beta_{j_{-i}} \sim N\left(\bar{\beta}_{j(i)}, \frac{\tau_{\beta_j}}{r_i}\right) \qquad \text{with} \qquad \bar{\beta}_{j(i)} = \sum_{k \in \partial_i} \beta_{j_k} / r_i$$

where $\beta_{j_{-i}}$ are the vectors of all β_j's excluding β_{j_i}; ∂_i is the set of neighbors of area i and r_i is the number of neighbors of area i.

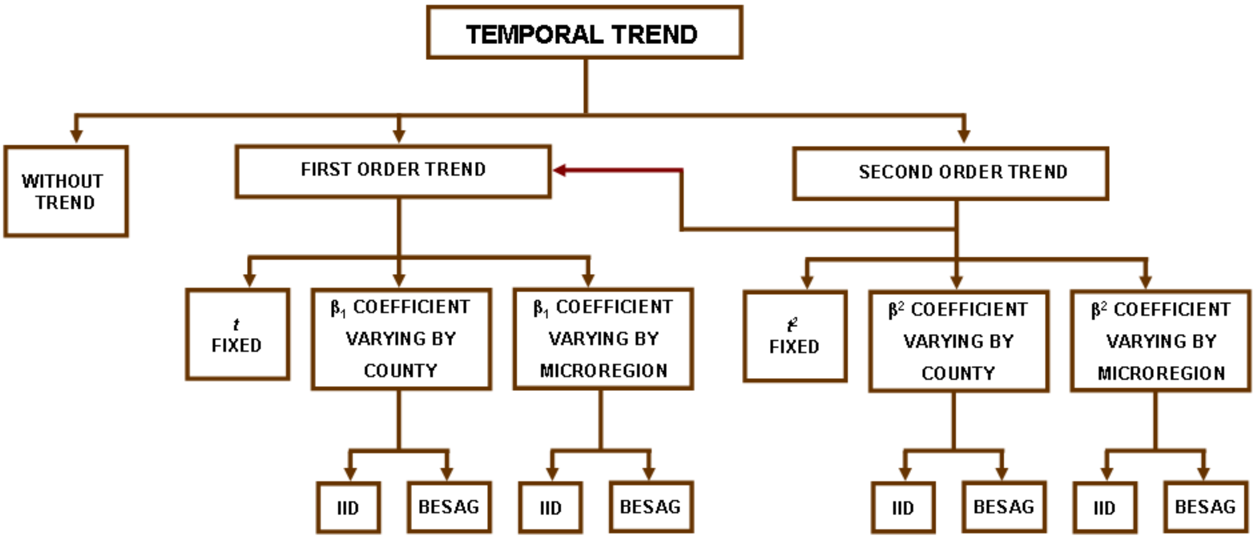
TEMPORAL TREND
WITHOUT TREND
FIRST ORDER TREND
SECOND ORDER TREND
t FIXED
β1 COEFFICIENT VARYING BY COUNTY
β1 COEFFICIENT VARYING BY MICROREGION
t² FIXED
β² COEFFICIENT VARYING BY COUNTY
β² COEFFICIENT VARYING BY MICROREGION
IID
BESAG
IID
BESAG
IID
BESAG
IID
BESAG

Figure 1. Schematic representation of the temporal trend variations.

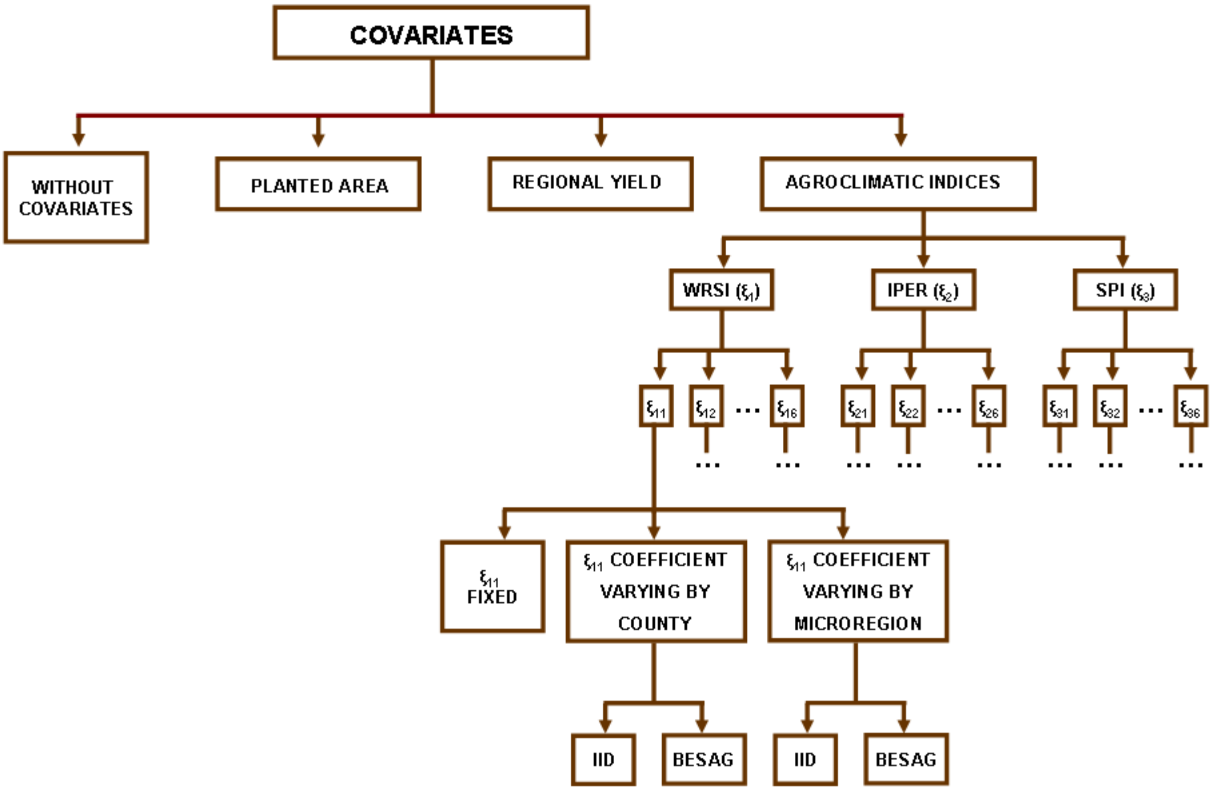
COVARIATES
WITHOUT COVARIATES
PLANTED AREA
REGIONAL YIELD
AGROCLIMATIC INDICES
WRSI (ξ_1)
IPER (ξ_2)
SPI (ξ_3)
ξ_{11} ξ_{12} ... ξ_{16}
ξ_{21} ξ_{22} ... ξ_{26}
ξ_{31} ξ_{32} ... ξ_{36}
ξ_{11} FIXED
ξ_{11} COEFFICIENT VARYING BY COUNTY
ξ_{11} COEFFICIENT VARYING BY MICROREGION
IID
BESAG
IID
BESAG

Figure 2. Schematic representation of the covariates variations.

3.2 Model selection criteria

We used criteria based on the posterior predictive distributions. As there is a two-year lag between the receipt of historical yield data and the deadline required for filling new contract terms, we consider the mean square predictive error at $T+1$ ($MSPE_1$) and at $T+2$ ($MSPE_2$) relative to the number of regions used in the analysis. Additionally, the deviance information criterion - DIC (Spiegelhalter et al., 2002) and predictive measures obtained from the INLA output (logarithmic score and the PIT histogram), were considered.

4 Main Results

As stated before, our main interest is on models with the best predictive behavior at $T+2$. the best models found at the first stage evaluation according to the model selection criteria are shown in Table 1. All of them include the autoregressive term, an intercept randomly varying by microregion, a temporal trend, with t's coefficients varying by county and t^2's coefficients varying by county or microregion (spatially or iid) and with the WRSI agroclimatic indexes randomly varying by county. Planted area was also present in the best models as a fixed covariate. Regional yield was not important at this stage.

Table 1. Best models found at the first stage evaluation.

β_0	β_1	β_2	ξ_1	$MSPE_1$ $\times 10^3$	$MSPE_2$ $\times 10^3$	pD	DIC
IID(m)	IID(c)	IID(c)	IID(c)	1106.3	1067.9	970	−781
IID(m)	IID(c)	BESAG(c)	IID(c)	1007.7	1116.6	946	−888
IID(m)	IID(c)	IID(m)	IID(c)	1070.5	1120.8	888	−854
IID(m)	BESAG(c)	IID(c)	IID(c)	1039.1	1122.4	859	−984
IID(m)	BESAG(c)	IID(m)	IID(c)	985.4	1160.7	764	−1060
IID(m)	BESAG(c)	BESAG(c)	IID(c)	939.3	1201.3	798	−1122
IID(m)	IID(m)	IID(m)	IID(c)	1085.5	1205.5	584	−854

(c): coefficients varying by county; *(m):* coefficients varying by microregion; ξ_1: WRSI covariate with its 6 subperiods included at a time in the model; *pD:* effective number of parameters.

Models in Table 1 were fitted again considering now all combinations of critical subperiods for the WRSI covariate. Best results are shown in Table 2.

Table 2. Best models found at the second stage evaluation.

ξ_{11}	ξ_{12}	ξ_{13}	ξ_{14}	ξ_{15}	ξ_{16}	$MSPE_1$ $\times 10^3$	$MSPE_2$ $\times 10^3$	pD	DIC
	✓		✓	✓		1016.3	952.3	847	−342
	✓		✓	✓	✓	1011.0	959.4	880	−663
	✓		✓			1014.8	959.6	827	−254
	✓		✓		✓	1011.4	973.9	873	−626
	✓				✓	945.8	986.4	772	154
			✓	✓	✓	1005.2	988.8	696	343
	✓			✓	✓	945.1	992.7	866	96

ξ_{11} to ξ_{16} correspond to the WRSI covariates for subperiods 1 to 6 respectively; *pD:* effective number of parameters.

The best models found at the second stage evaluation are shown in Table 2. All of them include the autoregressive term, planted area as a fixed covariate, an intercept randomly varying by microregion, a temporal trend, with t's coefficients spatially varying by county and t^2's coefficients randomly varying by microregion. The WRSI agroclimatic indexes associated to subperiods 1 and 3 did not contribute to decrease the MSPE values at $T+1$ and $T+2$. In contrast, the WRSI indexes associated to combinations of subperiods 2, 4, 5 and 6 randomly varying by county were very important to improve the model predictions. An example of how to specify the model formula to be read by the INLA library and the call to fit the best model in table 2 within R is given in the appendix.

The above results pointed out INLA as a flexible tool appropriate for fitting and compare a huge number of spatio-temporal crop yield models in an efficient way. Premium rates can be derived directly from the predictions of the best Bayesian hierarchical models identified. Moreover, the Bayesian approach allows one to derive the standard error estimates of the premium rates.

5 Extensions and Future work

The above methodology is currently being extended to a dynamic modelling setting, considering the spatio-temporal dynamic models for Gaussian areal data proposed in Vivar and Ferreira (2009). To illustrate the proposed extension we consider a non-stationary second-order spatio-temporal dynamic model. Therefore, for each time t and area s, $t = 1, \cdots, T$; $s = 1, \cdots, S$, we have that the observational and system equations of the dynamic model are given by

$$\boldsymbol{y}_t = \boldsymbol{F}_t' \boldsymbol{x}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim PGMRF\left(\boldsymbol{0s}, \boldsymbol{V}^{-1}\right) \tag{3}$$

$$\boldsymbol{x}_t = \boldsymbol{G}_t \boldsymbol{x}_{t-1} + \boldsymbol{\omega}_t, \qquad \boldsymbol{\omega}_t \sim PGMRF\left(\boldsymbol{0s}, \boldsymbol{W}^{-1}\right) \tag{4}$$

where,

$$\boldsymbol{x}_t = \begin{pmatrix} \boldsymbol{x}_{1t} \\ \boldsymbol{x}_{2t} \end{pmatrix}, \quad \boldsymbol{F}_t = \begin{pmatrix} \boldsymbol{Is} \\ \boldsymbol{0s} \end{pmatrix}, \quad \boldsymbol{G}_t = \begin{pmatrix} \rho_1 \boldsymbol{Is} & \rho_1 \boldsymbol{Is} \\ \boldsymbol{0s} & \rho_2 \boldsymbol{Is} \end{pmatrix} \text{ and } \boldsymbol{W}^{-1} = \begin{pmatrix} \boldsymbol{W}_1^{-1} & \boldsymbol{0s} \\ \boldsymbol{0s} & \boldsymbol{W}_2^{-1} \end{pmatrix}.$$

Here $\boldsymbol{y}_t = (y_{t1}, \cdots, y_{tS})'$ denote the observed field at time t; $\boldsymbol{0s}$ is a $S \times S$ null matrix, and $\boldsymbol{Is}$ is the $S \times S$ identity matrix. This model has two state fields: $\boldsymbol{x}_{1t}$ represents the level and $\boldsymbol{x}_{2t}$ represents the velocity of the process at time t. Moreover, we have

$$\boldsymbol{y}_t = \boldsymbol{x}_{1t} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim PGMRF\left(\boldsymbol{0s}, \boldsymbol{V}^{-1}\right), \tag{5}$$

$$\boldsymbol{x}_{1t} = \boldsymbol{x}_{1,t-1} + \boldsymbol{x}_{2,t-1} + \boldsymbol{\omega}_{1_t}, \qquad \boldsymbol{\omega}_{1_t} \sim PGMRF\left(\boldsymbol{0s}, \boldsymbol{W}_1^{-1}\right), \tag{6}$$

$$\boldsymbol{x}_{2t} = \boldsymbol{x}_{2,t-1} + \boldsymbol{\omega}_{2_t}, \qquad \boldsymbol{\omega}_{2_t} \sim PGMRF\left(\boldsymbol{0s}, \boldsymbol{W}_2^{-1}\right), \tag{7}$$

where the errors $\boldsymbol{\varepsilon}_t = (\varepsilon_{t1}, \cdots, \varepsilon_{tS})$ and $\boldsymbol{\omega}_{i_t} = (\omega_{i_{t1}}, \ldots, \omega_{i_{tS}})$, $i = 1, 2$, are independent and modeled as proper Gaussian Markov Random Fields (PGMRF). Matrices $\boldsymbol{W}_i$ and $\boldsymbol{V}$ describe the spatial covariance structure of $\boldsymbol{\omega}_{i_t}$ and $\boldsymbol{\varepsilon}_t$ respectively. Following Vivar and Ferreira (2009), precision matrices $\boldsymbol{W}_i$ and $\boldsymbol{V}$ are modelled as $\boldsymbol{W}_i = \tau_{\omega_i}(\boldsymbol{Is} + \phi_{\omega_i}\boldsymbol{M})$ and $\boldsymbol{V} = \tau_v(\boldsymbol{Is} + \phi_v \boldsymbol{M})$, with

$$M_{k,j} = \begin{cases} m_k & \text{if} \quad k = j, \\ -h_{k,j} & \text{if} \quad k \in N_j, \\ 0 & \text{otherwise.} \end{cases}$$

N_j is the set of neighbours of region j, $h_{k,j} > 0$ is a measure of similarity between regions k and j (here we assume that $h_{k,j} = 1$), $m_k = \sum_{j \in N_k} h_{k,j}$. τ_{ω_1}, τ_{ω_2} and τ_v are scale parameters and $\phi_{\omega_1}, \phi_{\omega_2}, \phi_v \geq 0$ control the degree of spatial correlation.

State space models (also known as *dynamic models* in the Bayesian literature) are not currently in the list of latent models fitted by INLA. However, it is possible to formulate a specific latent model in a state-space form in order to perform approximate inference on it using INLA. We propose here an approach that enable us to use the INLA library to perform inference in dynamic linear models. The approach consists in merging the actual observations from the observational equation with "pseudo" observations coming from the system equations of the state space model in a unique structure and fit this augmented latent model in INLA considering different likelihoods for the states and observations.

The "trick" to fit this model with INLA consists in equating to zero the system equations, that is, we re-write (6) and (7) as

$$\boldsymbol{0} = \boldsymbol{x}_{1t} - \boldsymbol{x}_{1,t-1} + \boldsymbol{x}_{2,t-1} + \boldsymbol{\omega}_{1_t}, \qquad \boldsymbol{\omega}_{1_t} \sim PGMRF\left(\boldsymbol{0s}, \boldsymbol{W}_1^{-1}\right),$$

$$\boldsymbol{0} = \boldsymbol{x}_{2t} - \boldsymbol{x}_{2,t-1} + \boldsymbol{\omega}_{2_t}, \qquad \boldsymbol{\omega}_{2_t} \sim PGMRF\left(\boldsymbol{0s}, \boldsymbol{W}_2^{-1}\right)$$

and then we build an augmented model with dimension $S \times T + 2(S \times T - S)$ merging these “faked zero observations” from the system equation with the actual observations from the observational equation in a unique structure, as shown in Figure 5, where the first $S \times T$ elements correspond to the stacked vector of (actual) observations, $\boldsymbol{y} = \{y_{11}, \cdots, y_{ST}\}$, while the remaining $2(S \times T - S)$ elements, corresponding to the number of state parameters in equations (6) and (7), are forced to be zero.

$$\left[\begin{array}{c:c:c}
y_{11} & \text{NA} & \text{NA} \\
\vdots & \vdots & \vdots \\
y_{ST} & \text{NA} & \text{NA} \\
\hdashline
\text{NA} & 0 & \text{NA} \\
\vdots & \vdots & \vdots \\
\text{NA} & 0 & \text{NA} \\
\hdashline
\text{NA} & \text{NA} & 0 \\
\vdots & \vdots & \vdots \\
\text{NA} & \text{NA} & 0
\end{array}\right]
\begin{array}{l}
\\ \\ \\
\left.\begin{array}{c} \\ \\ \\ \end{array}\right\} \text{(ST - S) elements} \\
\left.\begin{array}{c} \\ \\ \\ \end{array}\right\} \text{(ST - S) elements}
\end{array}$$

Figure 5. Schematic representation of the data structure for the augmented model.

Inference in this augmented model using the INLA approach is performed now considering three different likelihoods. The first $S \times T$ data points are assumed to follow a Gaussian distribution with unknown precision matrix $\boldsymbol{V}^{-1}$, whereas the last $2(S \times T - 1)$ data points, which are faked to be 0, are considered as observed with a high and fixed precision.

A simulated data set was generated in order to illustrate the application of the above method. The simulated data consisted of a time series of 30 times for each of the 100 areas of the North Carolina's map (that is, $S = 100$ and $T = 30$). This map is available in `R` from spdep package. Inference was performed for the state vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ as well as for the scale and correlation parameters, but not for the ρ's, whose values were fixed in one before analysis, leading to a non-stationary process. a summary of the posterior density for the hyperparameters is shown in Figure 6. True values felt into the 95% credibility intervals in all cases.

```
                   0.025quant     0.5quant  0.975quant
\tau_v             23.8774631   28.6585845  34.3797637
\phi_v              0.6416379    0.7842687   0.8864998
\tau_w1            24.2701416   36.8375799  54.5635119
\phi_w1             0.8854153    0.9483026   0.9751655
\tau_w2            34.3423897   43.3199276  60.0260982
\phi_w2             0.6524924    0.7887360   0.9004694
```

Figure 6. Summary of the posterior for the hyperparameters. True simulated values were: $\tau_v = 30$, $\tau_{w_1} = 50$, $\tau_{w_2} = 50$, $\phi_v = 0.8$, $\phi_{w_1} = 0.9$ and $\phi_{w_2} = 0.9$.

States also were well recovered, as can be seen in Figure 7.

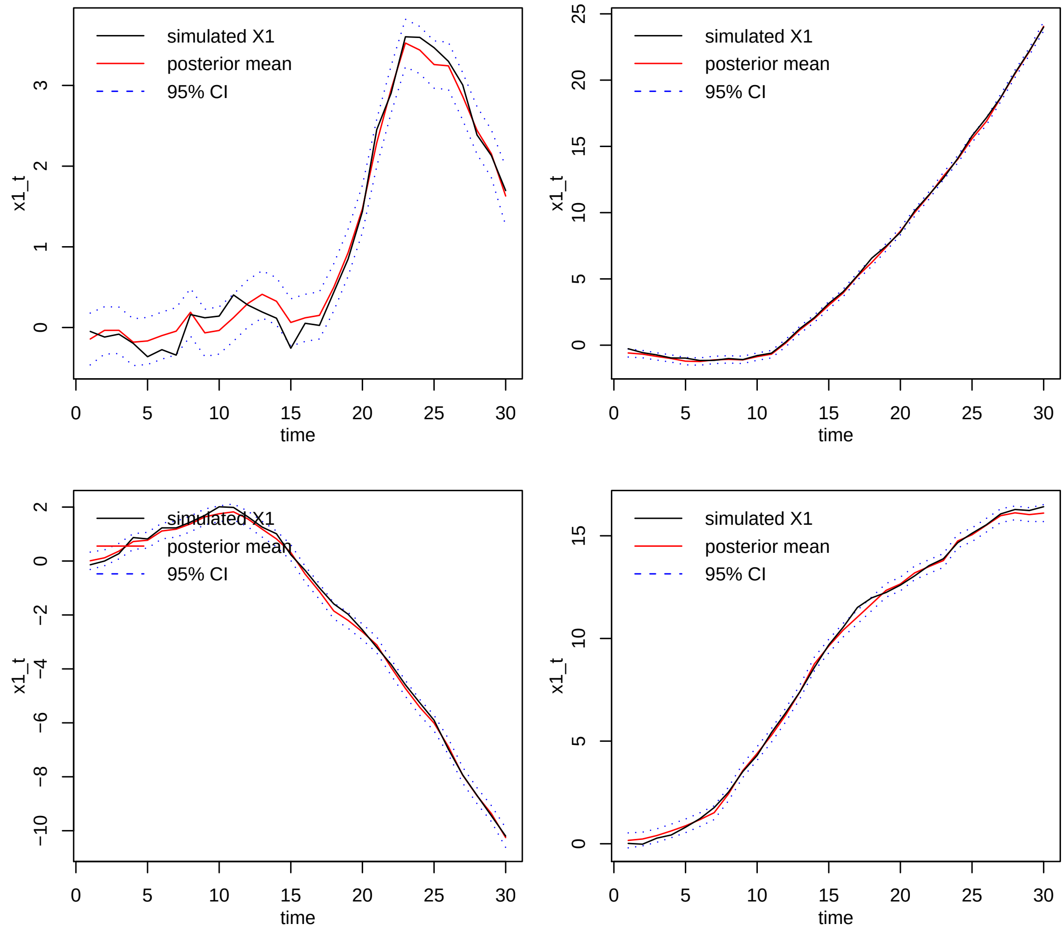

Figure 7. Simulated and predicted values (posterior mean and 95% credibility interval) for the the state vector X_1 at area 20 and its neighbors.

Extensions of this approach to include covariates and to forecast crop yields is straightforward and is currently being developed.

References

Blain, G.C. and Brunini, O. (2006) Quantificação da seca agrícola pelo índice padronizado de evapotranspiração real (IPER) no estado de São Paulo. *Bragantia*, **65**, 517–525.

Just, R.E., and Weninger, Q. (1999) Are CropYields Normally Distributed?. *American Journal of Agricultural Economics*, **81**, 287–304.

McKee, T.B., Doesken, N.J. and Kleist, J. (1993) The relationship of drought frequency and duration to time scales. In: Eight conference on Applied Climatology. Boston:American Meteorolgical Society, p.179–184.

Ozaki, V.A., Ghosh, S.K., Goodwin, B.K. and Shirota, R. (2008) Spatio-Temporal Modeling of Agricultural Yield Data with an Application to Pricing Crop Insurance Contracts. *American Journal of Agricultural Economics*, **90**, 951–961.

Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*. London: Chapman and HallCRC Press.

Rue, H., Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society series B*, **71**, 319–392.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society series B*, **64**, 583–639.

R Development Core Team (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Schrödle, B. and Held, L. (2009) Spatio-temporal disease mapping using INLA. Technical report, Biostatistics Unit, University of Zurich. Available at: <http://www.biostat.uzh.ch/research/manuscripts/schroedle_held_2009_3.pdf>

Vivar, J. C. and Ferreira, M. A. R. (2009) Spatiotemporal models for Gaussian areal data. *Journal of Computational and Graphical Statistics*, **18**, 658–674.

A Appendix

```
formula <- log(Y) ~ y.copy + areapl +
                    f(micro.b0, model='iid') +
                    time + f(s.bt, time, model='besag', graph.file='gpr') +
                    time2 + f(micro.bt2, time2, model="iid") +
                    i2 + f(s.i.2, i2, model = "iid") +
                    i4 + f(s.i.4, i4, model = "iid") +
                    i5 + f(s.i.5, i5, model = "iid") +
                    i6 + f(s.i.6, i6, model = "iid")

fit <- inla(formula, control.inla=list(h=0.2, strategy="GAUSSIAN"),
             control.compute=list(dic=TRUE, cpo=TRUE, mlik=TRUE), data=dfinla,
             control.predictor=list(compute=TRUE, cdf=c(.025, .975)),
             family=rep("gaussian", 5),
             control.data=list(list(param = c(1,0.01)),
                               list(param = c(1,0.01)),
                               list(param = c(1,0.01)),
                               list(param = c(1,0.01)),
                               list(param = c(1,0.01))))
```