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 promising alternative to inference via MCMC in latent Gaussian models are the integrated nested Laplace approximations (INLA) (Rue et al., 2009). The methodology is particularly attractive if the latent Gaussian model is a Gaussian Markov random field (GMRF). In this work, using the INLA approach, several spatio-temporal crop yield models were fitted and compared in an efficient way in order to identify the most suitable one to calculate the premium rate of an areal crop yield insurance contract for maize in Paraná state (Brazil).

2 Methodology

2.1 Hierarchical space-time crop yield models

The choice of a statistical model that adequately reflects the conditional density of yields is an important consideration in the actuarial calculation of an accurate premium rate. In doing this,

*Extended abstract presented at the 10^{th} Bayesian Statistics Brazilian Meeting – EBEB-X, Angra dos Reis, Brazil, March 21–24, 2010.

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

a number of issues relating to the modeling of crop yields must be considered, such as the fact that crop yields tend to have substantial trends and tend to be significantly correlated across space due to the systemic nature of weather.

In the construction of crop insurance contracts, it is typically the case that the terms and parameters of the contract must be available one to two years prior to the insurance cycle. This reflects the fact that crop yield data may take some time to be adequately measured following the harvest (Ozaki et al., 2008). Further, an insurance provider will not offer coverage after the insurance buyers have information about their yields. For example, contracts must typically be signed before planting season. If not, farmers may have an information advantage over insurers, who had to specify contract parameters at a much earlier date. In addition, administrative issues relating to the operation of any program require substantial lead time in providing the parameters of the contract offering. We will assume 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. 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 Bayesian inferential framework that accounts for all sources of uncertainty. As usually in 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.

A general version of the first kind of space-time models considered can be represented by the following hierarchical structure, where we simultaneously model the time trend and the temporal and spatial autocorrelation:

$$y_{it} \sim N(\mu_{it}, \tau_j), \qquad i \in \{1, \cdots, 397\}, \;\; t \in \{1, \cdots, 26\}, \;\; j \in \{1, \cdots, 5\}$$

$$\mu_{it} = \rho_i y_{i,t-1} + \beta_{0_i} + \beta_{1_i} t + \beta_{2_i} t^2 + \sum_{z=3}^{Z} \beta_{z_i} \, \xi_{zit}.$$

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 inverse gamma distributed 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 the 5 "macroregions" previously defined.

Several sub-models were fitted from this general model, considering different ways of entering with each of the terms and covariates in the models (see Figures 1 and 2). As covariates we consider the planted area in each county (in Hectares) and the aggregated crop yield at the "nucleo regional" level (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 (ISNA) and the Standardized Precipitation Index (SPI). The first two are based on ratios of actual and potential evapotranspiration 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 uses only precipitation data.

The determination of a unique critical period for maize (in terms of water requirements) to calculate the agroclimatic indices, which can be considered representative of each county is

difficult to establish since farmers in a same county sow different maize varieties in different dates. Instead, each of these indices were accumulated through 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 coefficients for t and t^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). Time 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:

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

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.

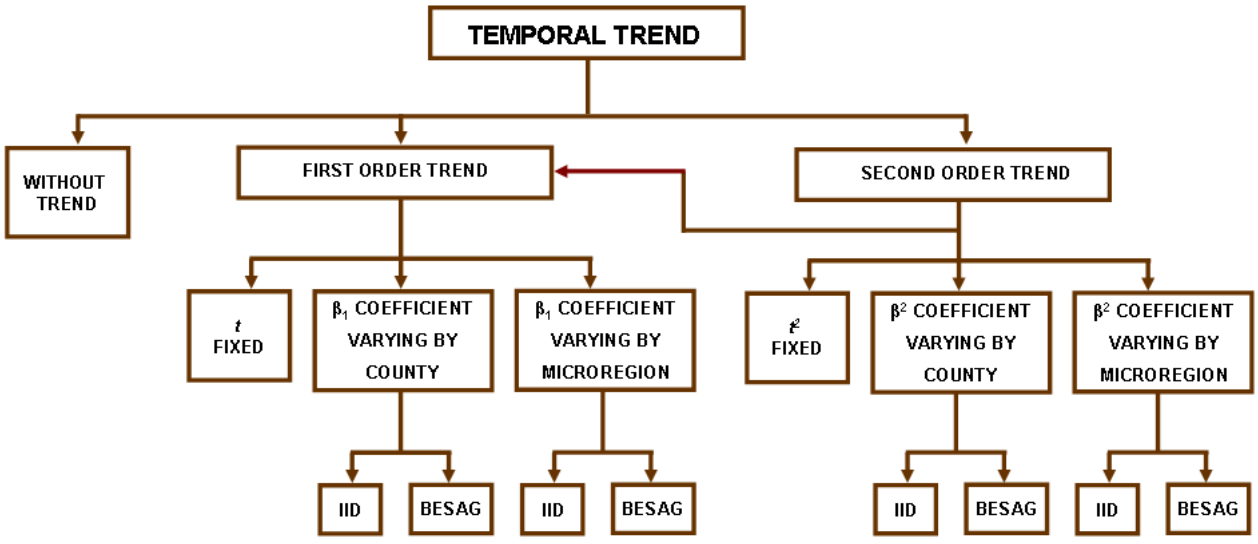

Figura 1. Schematic representation of the temporal trend variations.

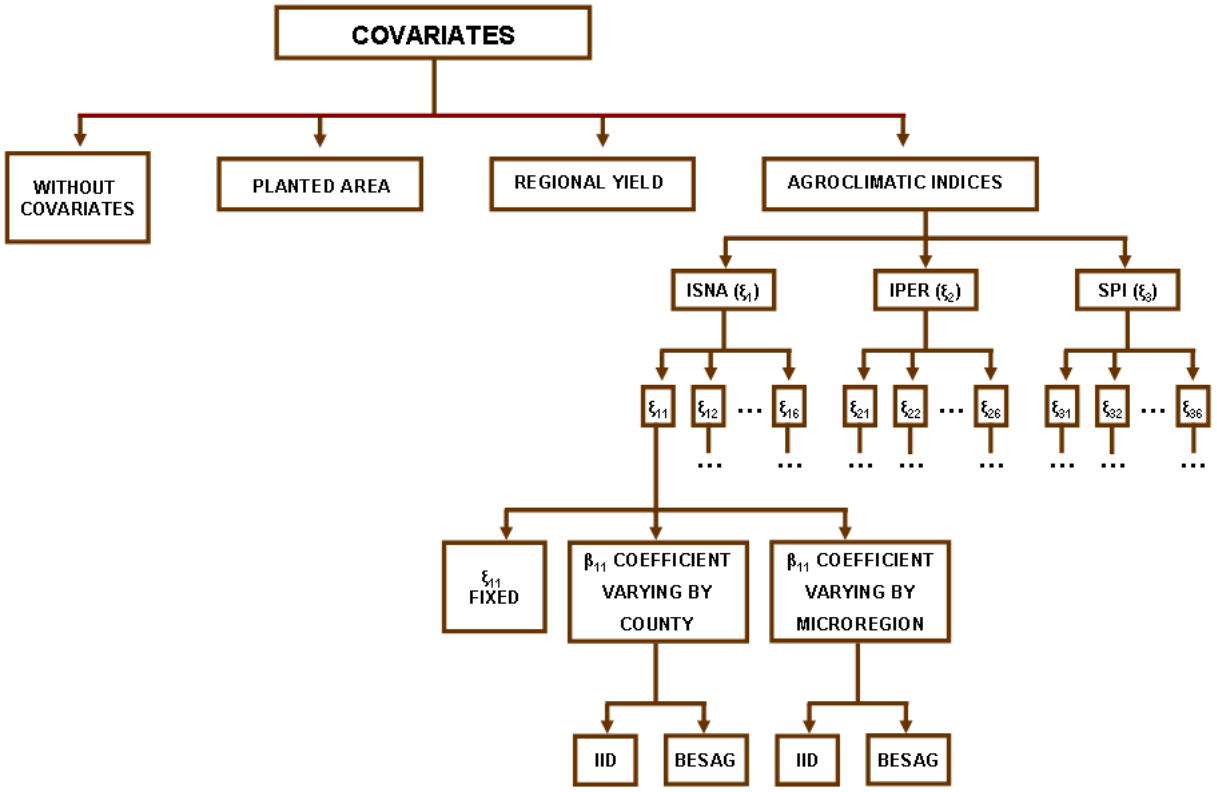

Figura 2. Schematic representation of the covariates variations.

2.2 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.

3 Integrated Nested Laplace Approximation (INLA)

The INLA computational approach combines Laplace approximations and numerical integration in a very efficient manner. In contrast with MCMC, the INLA method does not sample from the posterior. It approximates the posterior with a closed form expression. Therefore, problems of convergence and mixing are not an issue. The method is best suited to Bayesian hierarchical models for which there are 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).

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 $\pi(x_i \mid \boldsymbol{y})$ 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`.

4 Preliminary results

Best models for prediction at $T+2$ found at the first stage evaluation according to the model selection criteria are shown in Table 1. All of them include the autoregressive term, and 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 ISNA 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_1	β_2	ξ_1	$MSPE_1$	$MSPE_2$	pD	DIC
IID(c)	IID(c)	IID(c)	1106310	1067919	970	−781
IID(c)	BESAG(c)	IID(c)	1007761	1116614	946	−888
IID(c)	IID(m)	IID(c)	1070558	1120885	888	−854
BESAG(c)	IID(c)	IID(c)	1039166	1122424	859	−984
BESAG(c)	IID(m)	IID(c)	985481	1160771	764	−1060
BESAG(c)	BESAG(c)	IID(c)	939302	1201360	798	−1122
IID(m)	IID(m)	IID(c)	1085532	1205514	584	−854

(c): coefficients varying by county; *(m):* coefficients varying by microregion; ξ_1: ISNA 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 ISNA 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}	$MSPE_1$	$MSPE_2$	pD	DIC
	✓		✓	✓		1016348	952302	847	−342
	✓		✓	✓	✓	1011074	959412	880	−663
	✓		✓			1014841	959608	827	−254
	✓		✓		✓	1011425	973956	873	−626
	✓				✓	945832	986454	772	154
			✓	✓	✓	1005273	988888	696	343
	✓			✓	✓	945080	992760	866	96

ξ_{11} to ξ_{16} correspond to the ISNA 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 and a temporal trend, with t's coefficients spatially varying by county and t^2's coefficients randomly varying by microregion. The ISNA 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 ISNA 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 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.

5 Future work

We are currently exploring other models including space-time interactions and different ways of modelling the variance τ_i using INLA. We are also exploring the possibility of using the INLA approach for a similar study involving a spatio-temporal dynamic model for Gaussian areal data (Vivar and Ferreira, 2009).

References

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

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`.

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 + time +
                    f(s.bt, time, model='besag', graph.file='gpr') +
                    time2 + f(micro2, 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=c("gaussian", "gaussian", "gaussian", "gaussian", "gaussian"),
             control.data=list(list(param = c(1,0.1)),
                               list(param = c(1,0.1)),
                               list(param = c(1,0.1)),
                               list(param = c(1,0.1)),
                               list(param = c(1,0.1))))
```