Spatio-Temporal modelling for Pricing Area Yield Crop Insurance Contracts

Ramiro Ruiz Cárdenas (ramiro@est.ufmg.br)

Elias Teixeira Krainski (eliaskr@ufpr.br)

Collaborators:

Juan Carlos Vivar Rojas (Duke Univ.), **Vitor Augusto Ozaki** (ESALQ/USP)

Paulo Justiniano Ribeiro Jr (UFPR), **Håvard Rue** (NTNU), **Wagner Hugo Bonat** (UFPR)

IBC 2010 – Florianópolis, 09 December 2010

The Brazilian Agri-business

In 2009 it accounted for:

- 26.46% of the Gross Domestic Product
- 36.37% of Brazilian exports

Projections for 2010:

- harvesting of 135 million ton.

Brazil is the first export country of:

- coffee, sugar
- orange juice
- beef and chicken meat
- ethanol, tobacco and leather

It will soon be the first producer of:

- cotton
- soybean and vegetal oil
- bio-fuel made from sugarcane

Table: Evolution of the Brazilian crop insurance market

	2005	2006	2007	2008	2009
Subsidies (U$ million)	**1.3**	**16.8**	**33**	**84.9**	**147**
Policies (thousands)	**0.849**	**21.7**	**31.6**	**60**	> 100
Area (ha million)	**0.068**	**1.6**	**2.3**	**4.8**	> 8.1
Amount secured (U$ billion)	**0.07**	**1.6**	**1.5**	**3.9**	**8.3**

source: MAPA 2009

Adoption is still low ($< 10\%$).

Traditional crop insurance

Main Problems:

- **Lack of methods to properly quantify agricultural risk;**
- **Inadequate pricing techniques;**
- **Insufficient sources of data;**
- **Systemic nature of the risk;**
- **Information asymmetries (moral hazard, adverse selection);**
- **Greater ruin probability.**

Index based crop insurance contracts

Two main types:

- **Area yield insurance:**

  Farmers collect an indemnity whenever the “expected” area average yield (e.g. a county) falls beneath a yield guarantee, regardless of farmers’ actual yields.

- **Weather based insurance:**

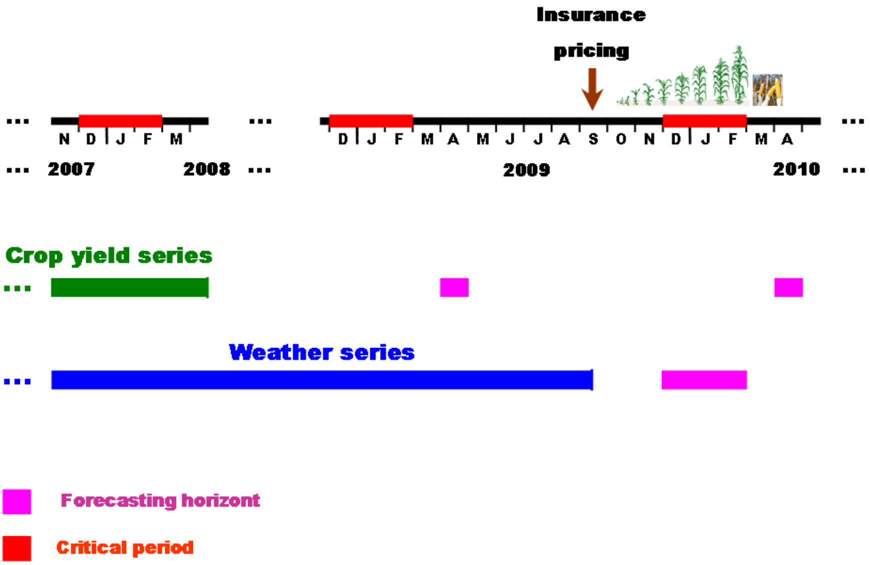
  It is based on the events of a weather variable measured at a given location. The payoff is based on the difference between an underlying weather index over a specified period and an agreed strike value.

Not currently implemented in Brazil.

Modelling issues

- **Large error for crop yield estimates in some regions;**
- **Short length of time series;**
- **Frequent missing values;**
- **Change of support issues;**
- **Variable sowing dates $\Rightarrow$ apparent weak correlation between covariates and yields;**
- **Two years delay in crop yield official statistics releases (practical and legal issues);**
- **Need of seasonal forecasting (4 months ahead) for agro-climatic covariates.**

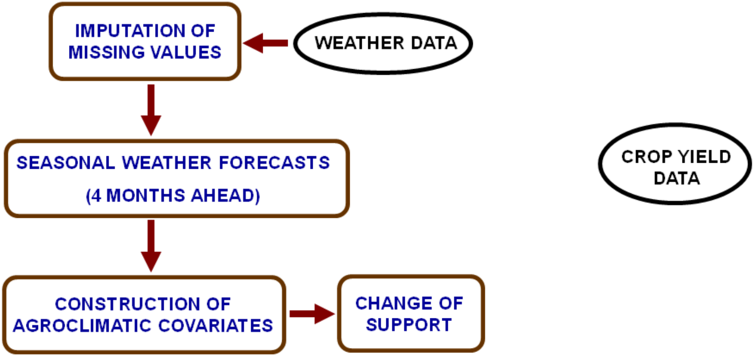
Chronological scheme



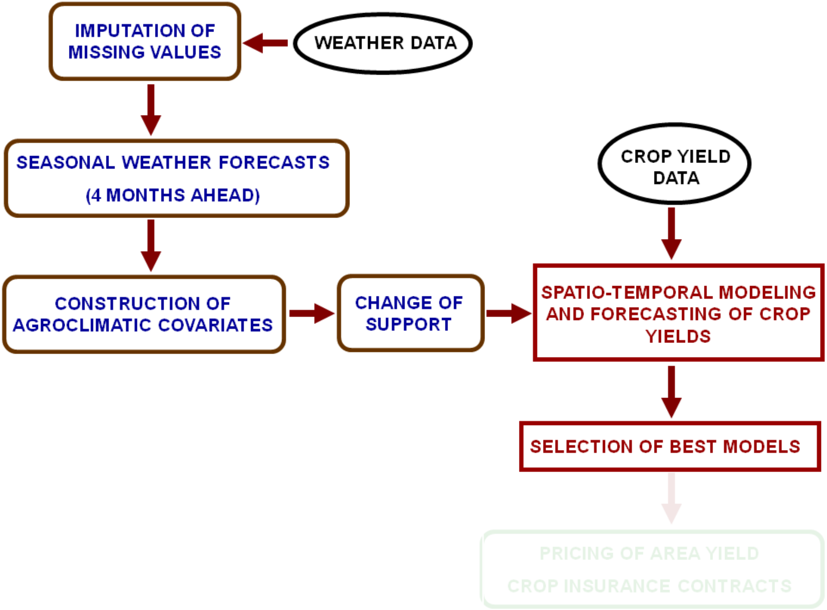
The proposed pricing approach

The proposed pricing approach

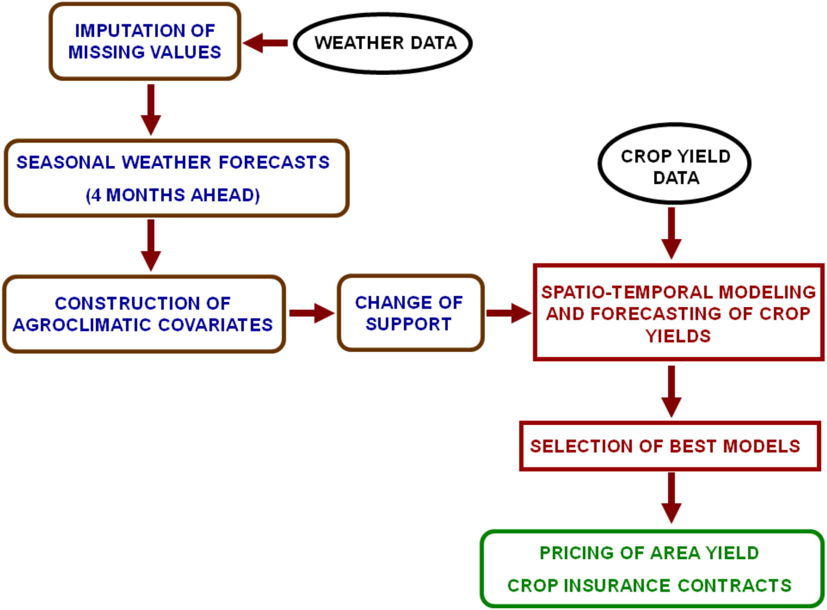
The proposed pricing approach



The proposed pricing approach



The proposed pricing approach



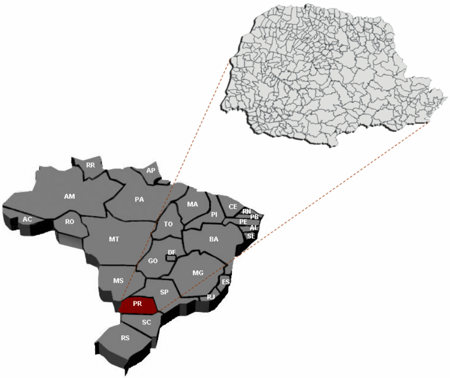
Study region and available data sets

- Crop yield data:
  average annual county corn yield (1980 – 2007).
  source: IBGE / SEAB

- Meteorological data:

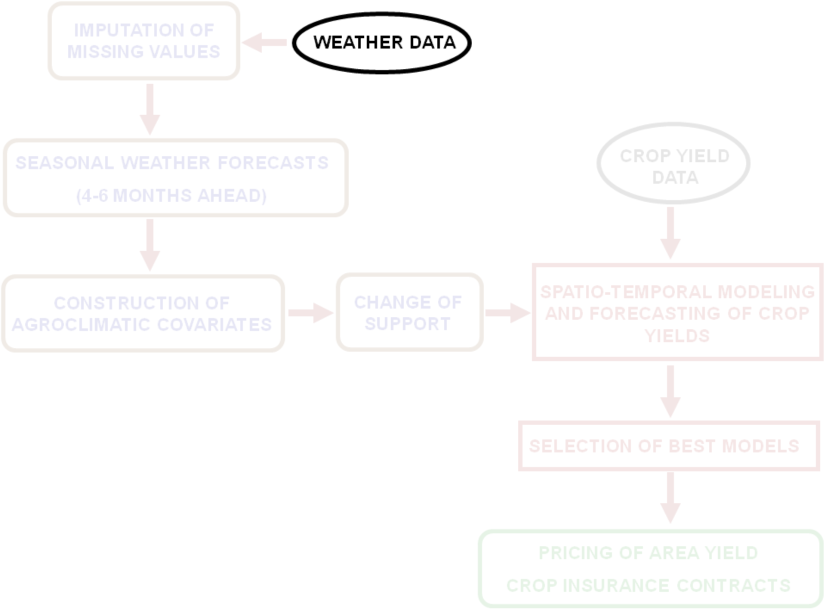
  daily precipitation for 503 stations (01/01/76 – 31/12/08).
  source: ANA / SUDHERSA / IAPAR / SIMEPAR / INMET

  daily temperature and solar radiation for 87 stations (01/01/76 – 31/12/08).
  source: INMET / IAPAR / SIMEPAR

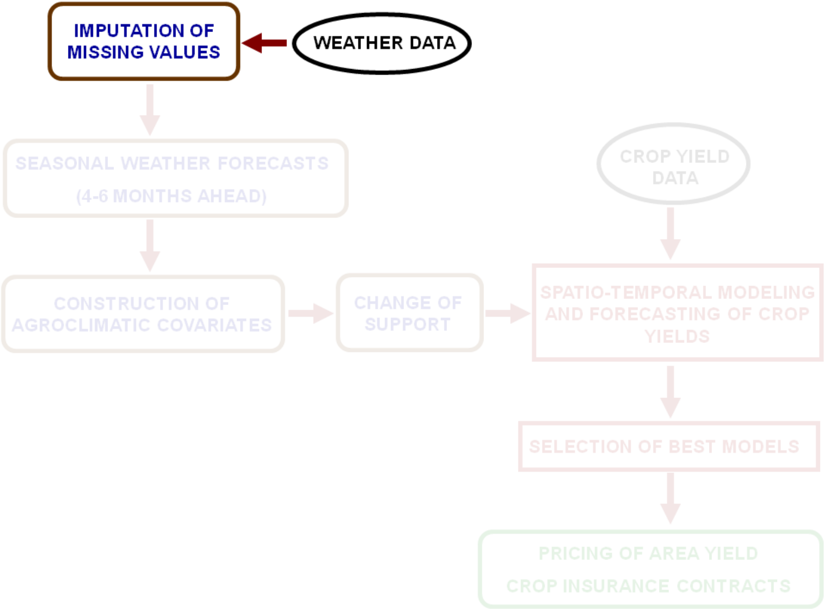


State: Paraná
Nº counties: 399
planted area (grains): 8.45 mil Ha

The proposed pricing approach

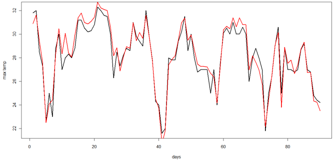


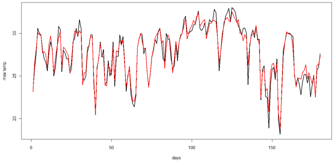
The proposed pricing approach

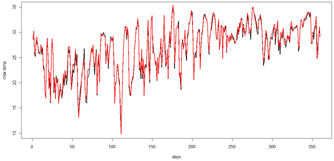


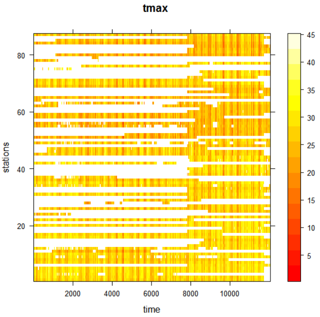
Imputation of temperature and solar radiation:

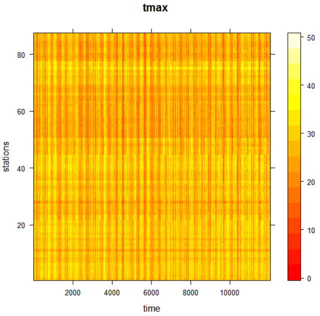
- We used a modified EM algorithm accounting for both spatial and temporal correlation structures *(Junger et al, 2003)*;









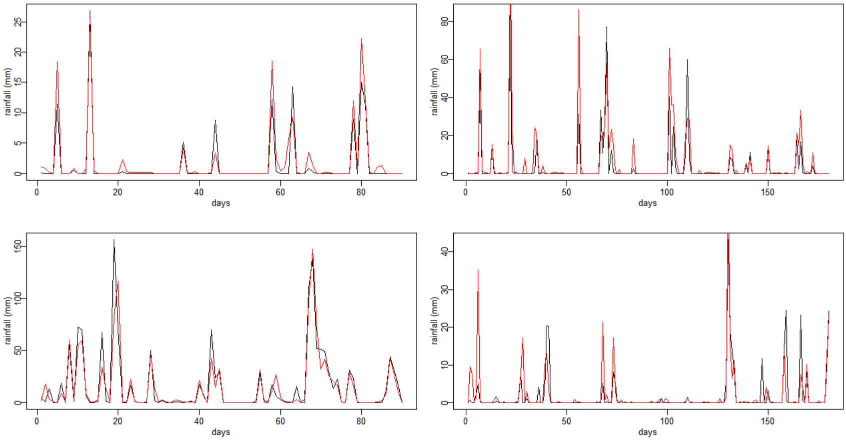


Imputing missing daily rainfall data:

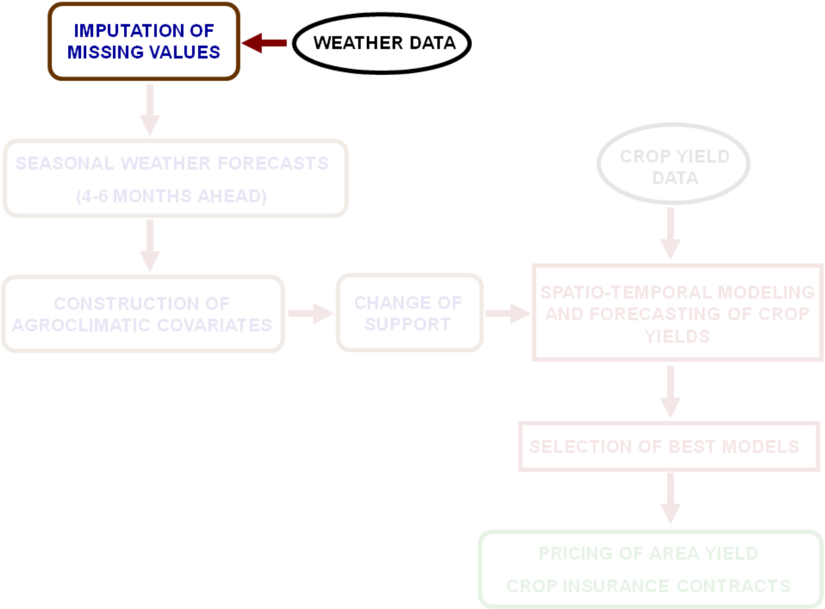
we used a probability density function-preserving approach *(Simolo et al., 2010)*

1. Occurrence is estimated by a weighting-based method modified by a wet/dry threshold;

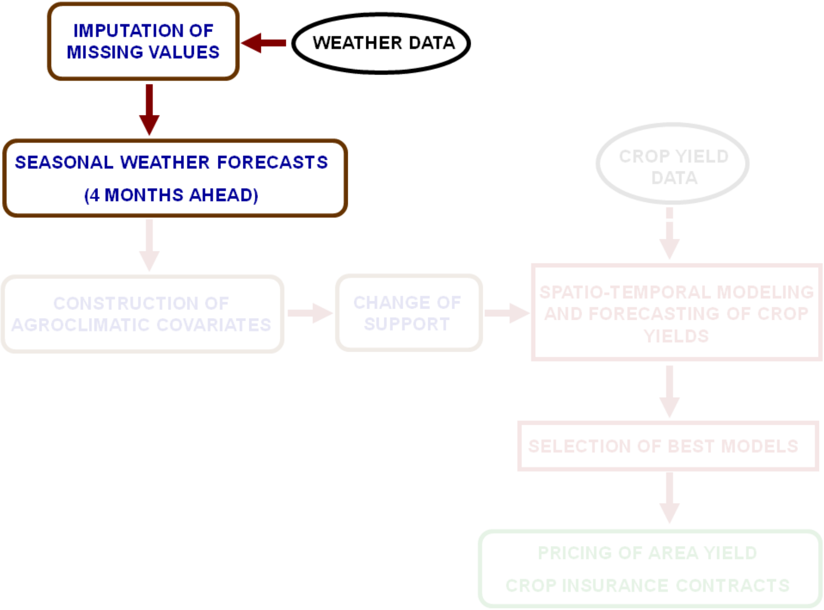
2. Full precipitation amount for wet-classified days is estimated by a modified multi-linear regression approach.



The proposed pricing approach



The proposed pricing approach



Seasonal forecast of weather variables:

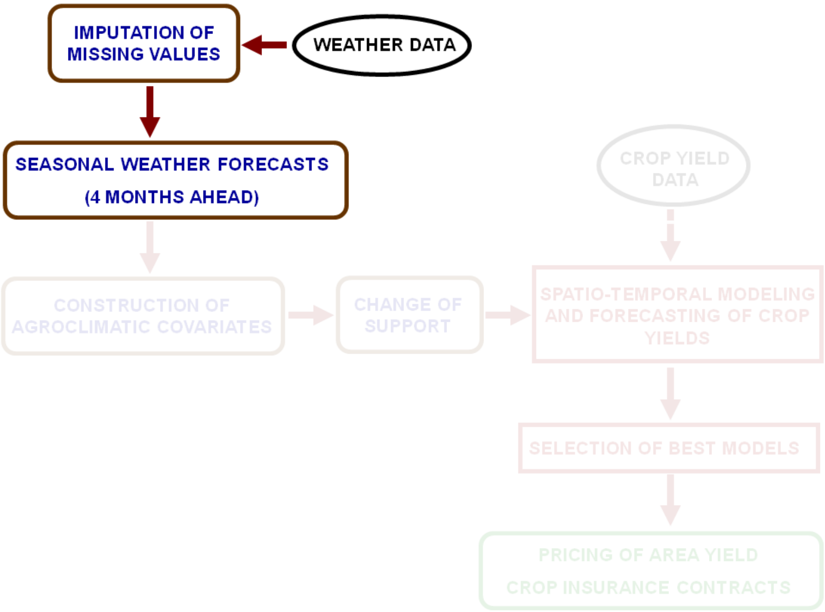
We are currently evaluating three main approaches:

- **Numerical forecasts from a regional model (Eta):**
  - easy to obtain,
  - high spatial resolution,
  - correction of systematic errors and temporal mismatch is needed.

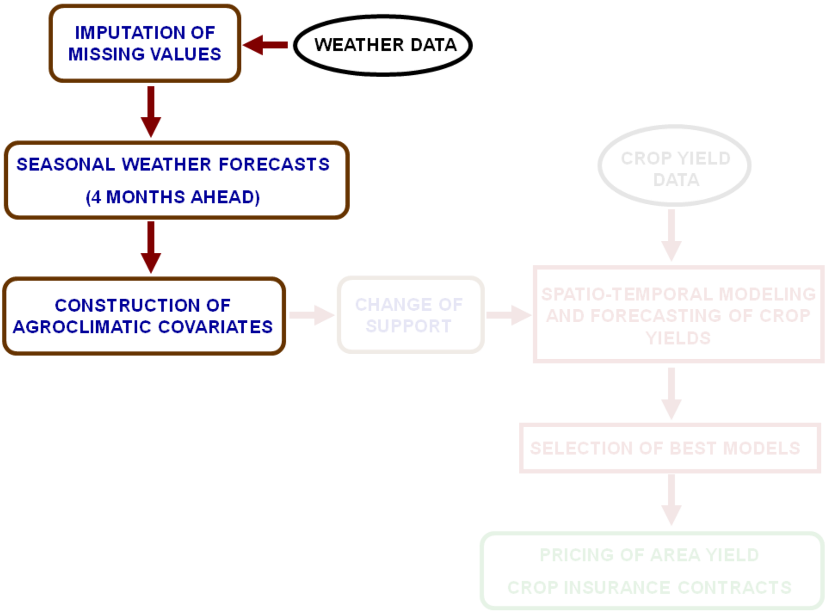
- **Weather analogue techniques:**
  - they recognize the most similar pattern, to the available data of the target year, among the same sequence of historical data.
  - Daily weather data of the selected year as the *best match* would be considered for the remainder of the target year.

- **Spatio-temporal dynamic models**
  - Daily rainfall:
    - A dynamic linear model based on a truncated normal distribution (Sansó and Guenni, 2000)
  - Daily temperature:
    - A dynamic seasonal regression (Huerta et al., 2004)

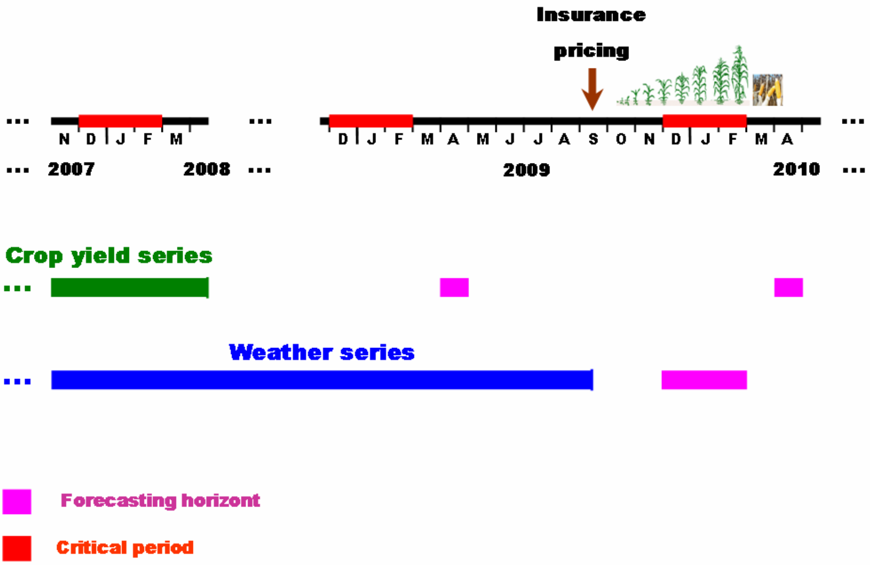
The proposed pricing approach



The proposed pricing approach



Chronological scheme



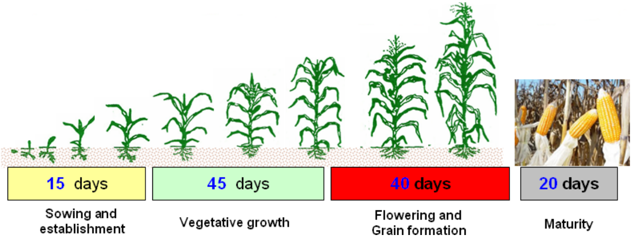
The agroclimatic Covariates:

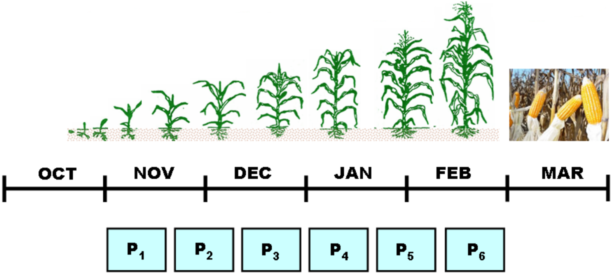
- *The Water Requirement Satisfaction Index [WRSI]:*
  It is based on the actual evapotranspiration to maximum evapotranspiration ratio (ETa/ETm) for a given "critical" period of the crop.

- *The standardized actual evapotranspiration index [IPER] (Blain et al., 2006):*
  It quantifies agricultural drought in a 10-days scale, based on the fit of the *ETa* series to the beta distribution.

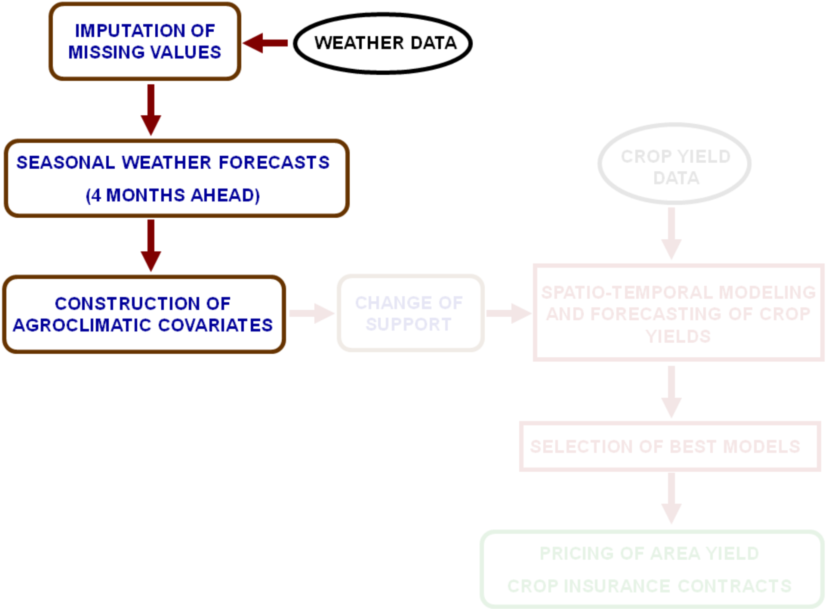
- *The Standardized Precipitation Index [SPI]* (McKee et al., 1993):
  It indicates the number of standard deviations that a particular precipitation event deviates from normal conditions.

Critical periods:

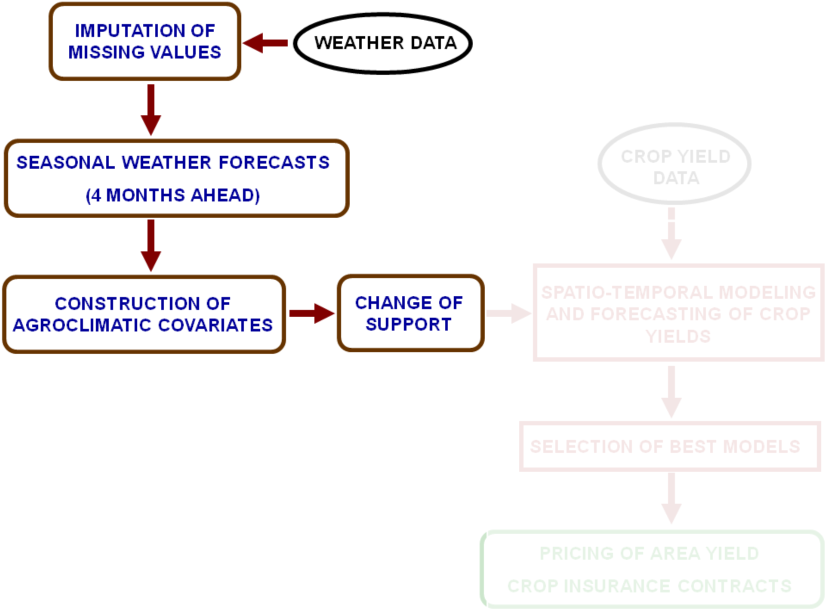




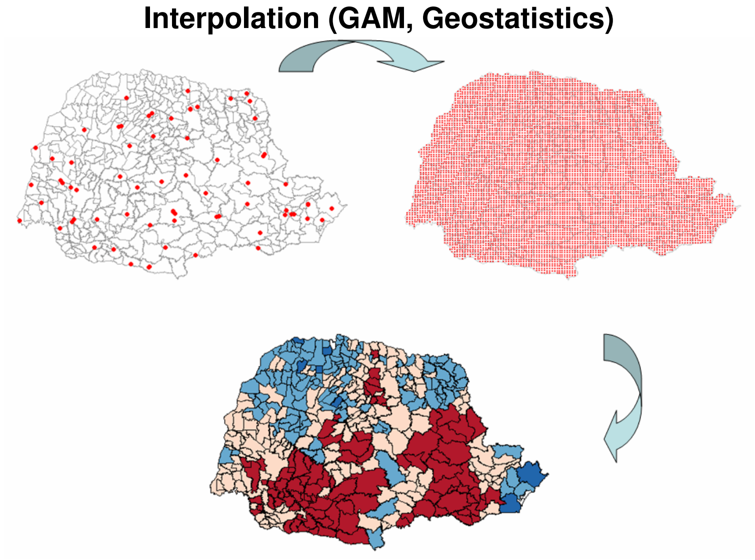
The proposed pricing approach



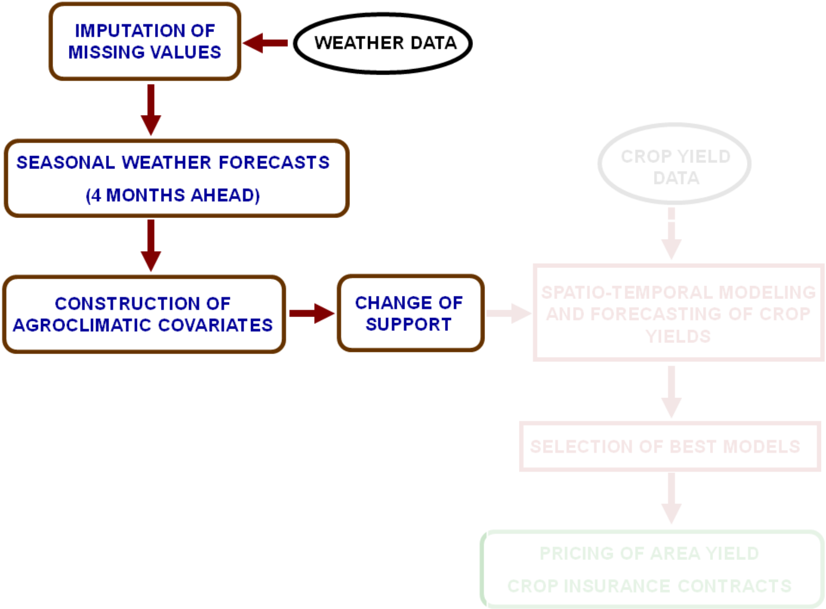
The proposed pricing approach



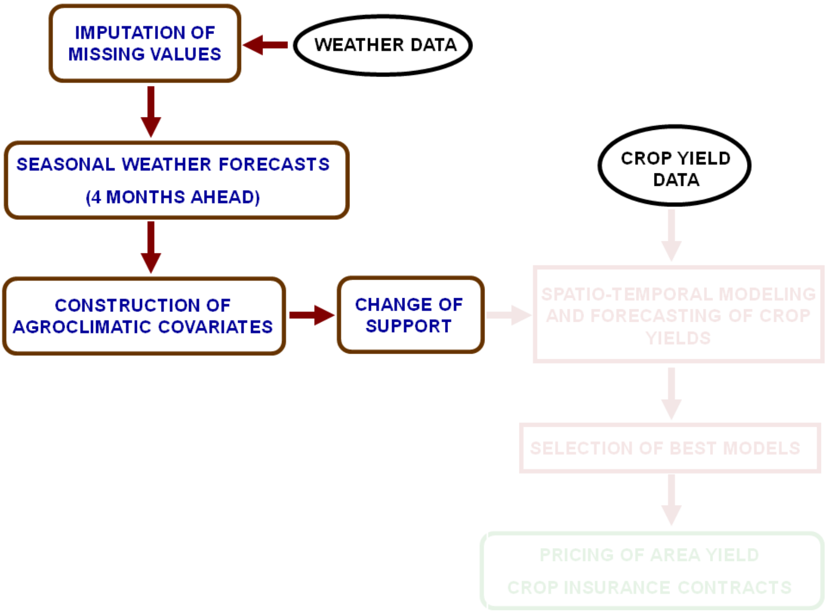
Solving the change of support problem



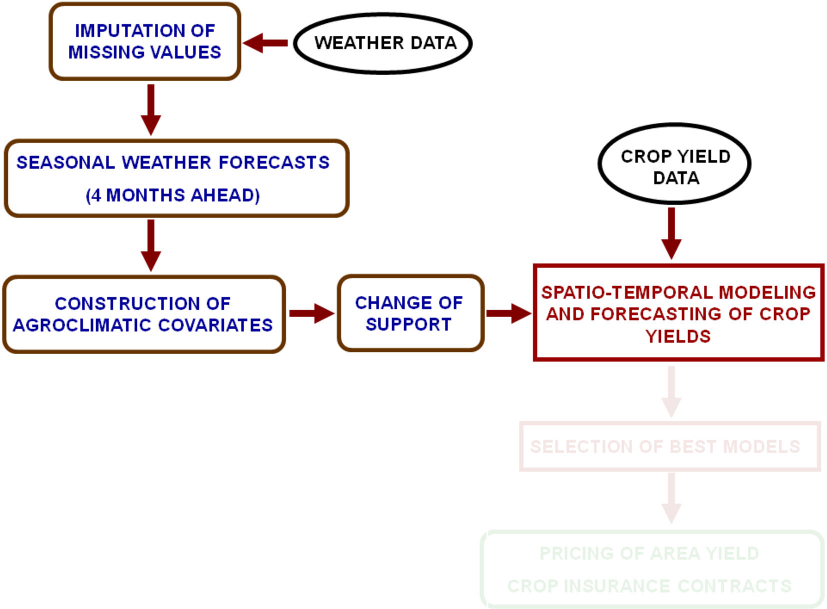
The proposed pricing approach



The proposed pricing approach



The proposed pricing approach



Spatio-temporal modelling of crop yields

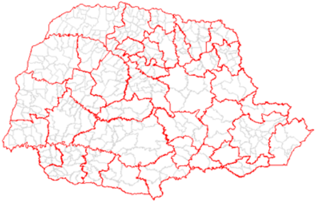
Bayesian approaches:

- A hierarchical space-time model

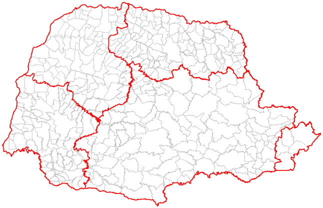
- A second order non-stationary spatio-temporal dynamic model

Levels of aggregation

County (399 areas)



Microregion (39 areas)



Region (5 areas)

The hierarchical model

$$\log(y_{st}) \sim N(\mu_{st}, \tau_r), \qquad i=1,\cdots,S; \quad t=1,\cdots,T; \quad r=1,\cdots,R; \quad R<S$$

$$\mu_{st} = \sum_{k=1}^{K} \rho_k y_{s,t-k} + \beta_{0_s} + \sum_{m=1}^{M} \beta_{m_s} t^m + \sum_{z=3}^{Z} \beta_{z_s}\, \xi_{zst}$$

$$\tau_r = \exp(\alpha_\tau + h_r),$$

where:

α_τ is the overall level of the log-volatility;

$$h_r \mid h_{-r} \sim N\left(\bar{h}_{(r)}, \frac{\tau_h}{\#g_r}\right) \qquad \text{with} \qquad \bar{h}_{(r)} = \sum_{p\in\partial_r} h_p/\#g_r$$

h_{-r} are the vectors of all h's excluding h_r;

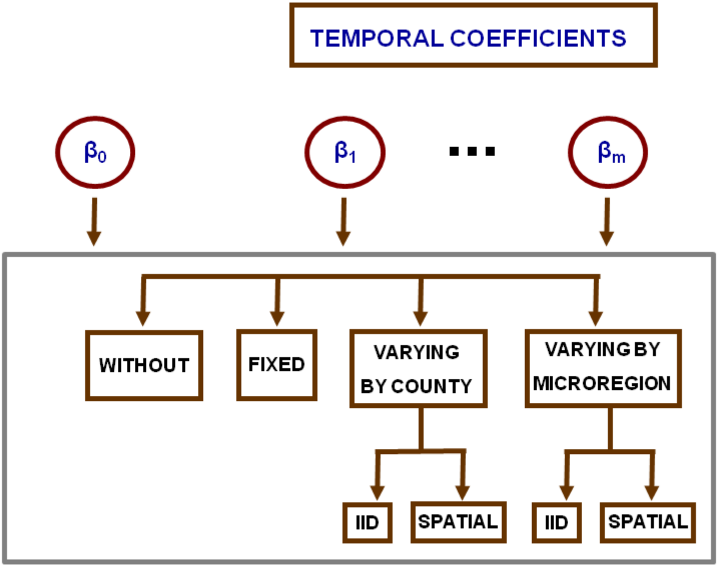
∂_r is the set of neighbors of region r;

$\#g_r$ is the number of neighbors of region r;

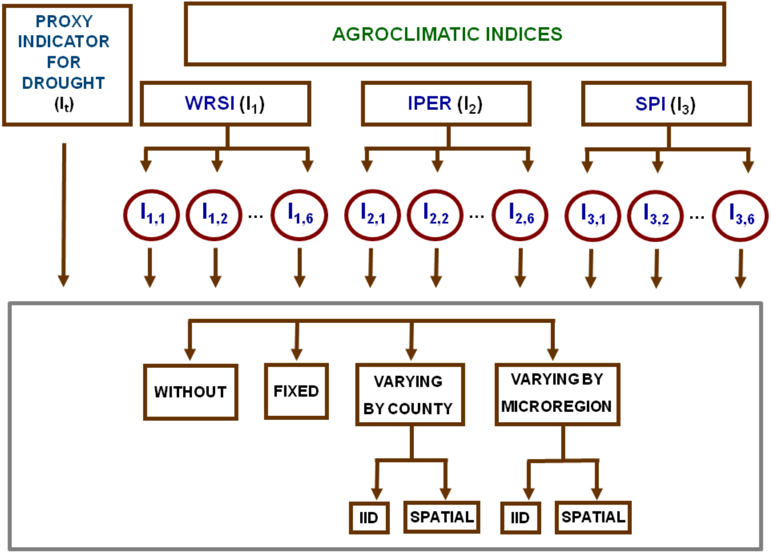
Schematic representation of the stochastic trend variation

- without a stochastic trend
- ρy_{t-1}
- $\rho_1 y_{t-1} + \rho_2 y_{t-2}$
- $\rho_1 y_{t-1} + \rho_2 y_{t-2} + \rho_3 y_{t-3}$
- $\rho \bar{y}^{(-2)} = \rho(y_{t-1} + y_{t-2})/2$
- $\rho \bar{y}^{(-3)} = \rho(y_{t-1} + y_{t-2} + y_{t-3})/3$
- $\rho \bar{y}^{(-4)} = \rho(y_{t-1} + y_{t-2} + y_{t-3} + y_{t-4})/4$

Schematic representation of the temporal trend variation



Schematic representation of the covariates variation



A Bayesian dynamic space-time model (Vivar and Ferreira, 2009)

$$y_t = x_{1t} + \Upsilon_{1t} z_{1t} + \cdots + \Upsilon_{6t} z_{6t} + \nu_t, \qquad \nu_t \sim PGMRF\left(0s, V^{-1}\right)$$

$$x_{1t} = x_{1,t-1} + x_{2,t-1} + \omega_{1_t}, \qquad \omega_{1_t} \sim PGMRF\left(0s, W_1^{-1}\right)$$

$$x_{2t} = x_{2,t-1} + \omega_{2_t}, \qquad \omega_{2_t} \sim PGMRF\left(0s, W_2^{-1}\right)$$

$$z_{1t} = z_{1,t-1} + \delta_{1_t}, \qquad \delta_{1_t} \sim PGMRF\left(0s, \Delta_1^{-1}\right)$$

$$\vdots \qquad\qquad \vdots$$

$$z_{6t} = z_{6,t-1} + \delta_{6_t}, \qquad \delta_{6_t} \sim PGMRF\left(0s, \Delta_6^{-1}\right)$$

where, for $i = 1, 2$, $j = 1, \ldots, 6$:

$$\Delta_j^{-1} = \tau_{\delta_j}(Is + \phi_{\delta_j} M) \qquad W_i^{-1} = \tau_{\omega_i}(Is + \phi_{\omega_i} M) \qquad V^{-1} = \tau_\nu(Is + \phi_\nu M)$$

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

$h_{k,\ell} > 0$ is a measure of similarity between regions

$\phi_{\delta_j}, \phi_{\omega_i}, \phi_\nu \geq 0$ control the degree of spatial correlation

N_ℓ is the set of neighbours of region ℓ

$\tau_{\delta_j}, \tau_{\omega_i}, \tau_\nu$ are scale parameters

$m_k = \sum_{\ell \in N_k} h_{k,\ell}$

For approximate Bayesian inference details see Ruiz-Cárdenas, Krainski and Rue (2010)

The need for a fast inference procedure

- A large number of models (over 10000) combining covariates, regional effects, time trends and space-time interactions must be fitted and compared;
- MCMC becomes unfeasible when the number of areas increases;

✓ **We use the INLA approach (Rue et al, 2009) to identify the best spatio-temporal models to use in the calculation of the premium rates of an area yield insurance contract for maize in Paraná state (Brazil)**

✓ INLA combines Laplace approximations and numerical integration to approximate posterior marginals;

✓ It is Best suited to Bayesian models with a large number of unknown parameters following a GMRF, $\pi(\boldsymbol{x} \mid \theta)$, and a small number of hyperparameters θ.

The need for a fast inference procedure

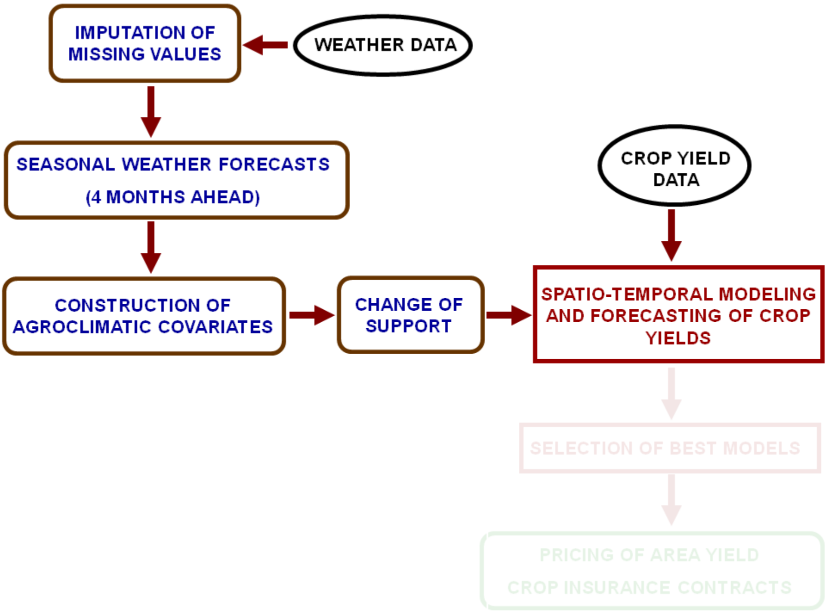
- A large number of models (over 10000) combining covariates, regional effects, time trends and space-time interactions must be fitted and compared;
- MCMC becomes unfeasible when the number of areas increases;

✓ **We use the INLA approach (Rue et al, 2009) to identify the best spatio-temporal models to use in the calculation of the premium rates of an area yield insurance contract for maize in Paraná state (Brazil)**

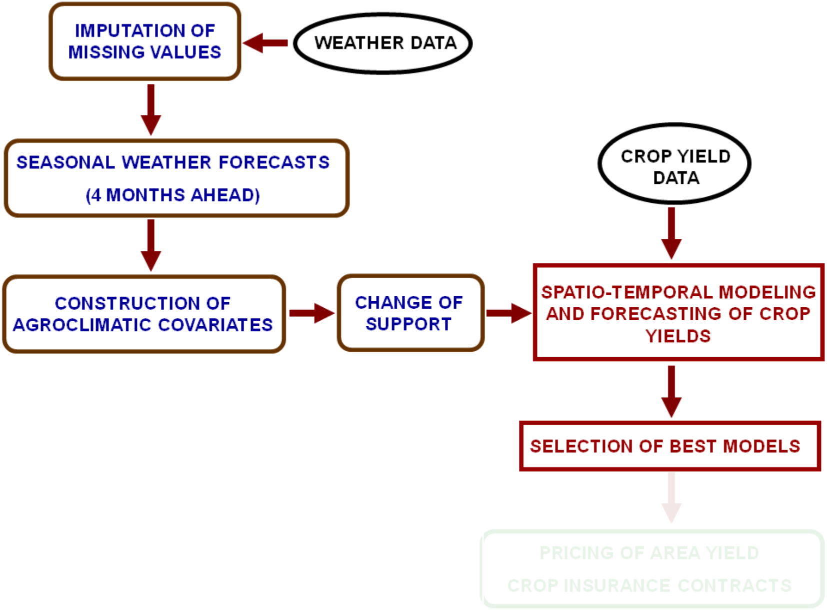
✓ INLA combines Laplace approximations and numerical integration to approximate posterior marginals;

✓ It is Best suited to Bayesian models with a large number of unknown parameters following a GMRF, $\pi(\boldsymbol{x} \mid \boldsymbol{\theta})$, and a small number of hyperparameters $\boldsymbol{\theta}$.

The proposed pricing approach



The proposed pricing approach



Model Selection

- **Forecast accuracy measures:**
  - weighted Mean Absolute Scaled Error (*wMASE*):

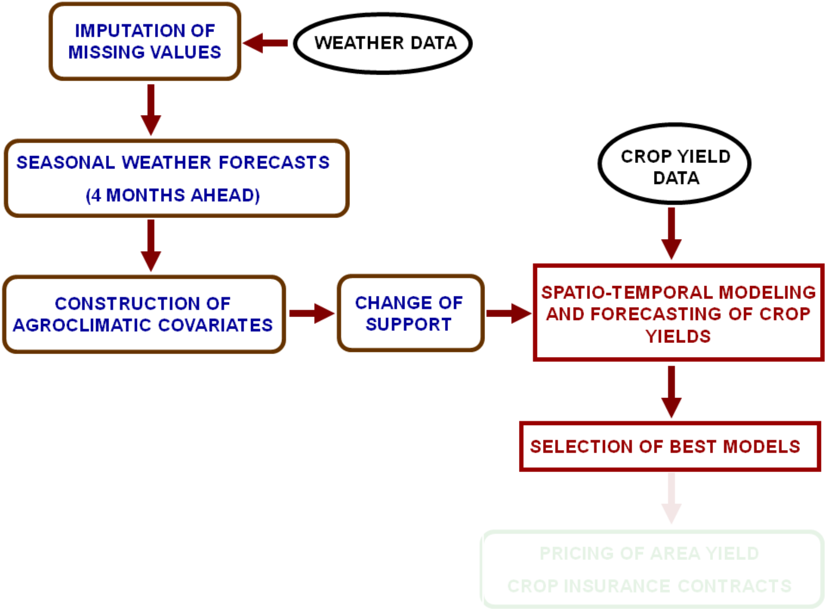
$$wMASE = \sum_{i=1}^{n} \left[\omega_{it} \frac{|y_{it} - \hat{y}_{it}|}{\frac{1}{n-1}\sum_{j=2}^{n} |y_{jt} - (y_{j,t-1} + (y_{j,t-1} - y_{j,t-2}))|} \right]$$

where $\omega_{it} = area_{it} / \sum_{j=1}^{n} area_{jt}$, $area_{it}$: planted area for county i at time t

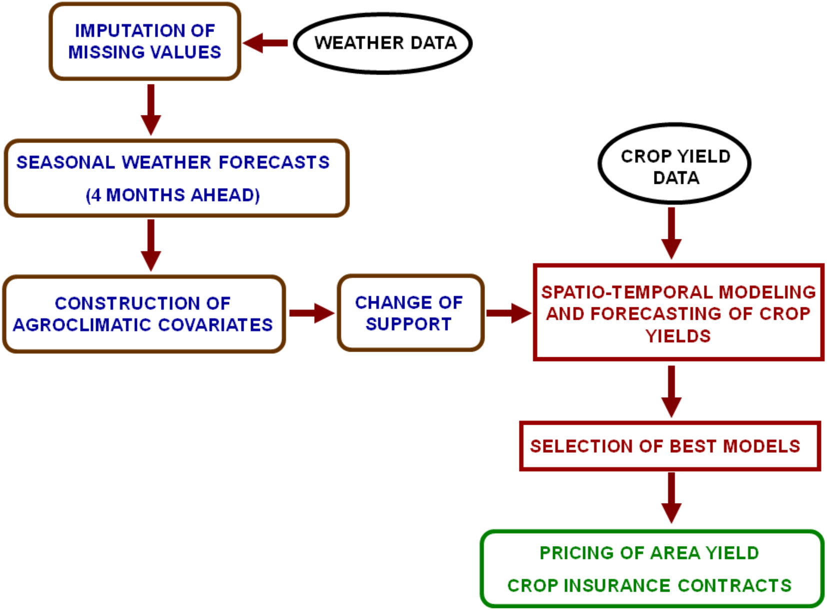
  - weighted Root Mean Squared Error (*wRMSE*),
  - weighted Mean Absolute Error (*wMAE*),
  - weighted Mean Absolute Percentage Error (*wMAPE*).

- Fitting was based on the first 26 years of data (last two years were leaved out)
- Two steps (years) ahead forecast
- Best models were those with better performance over all measures

The proposed pricing approach



The proposed pricing approach



Rating the crop insurance contract

The fair premium rate (PR):

$$\text{Indemnity} = \max\left\{0, \frac{\alpha y^e - y}{\alpha y^e}\right\}$$

$$\text{PR} = \frac{1}{\alpha y^e} \int_0^{\alpha y^e} (\alpha y^e - y) f(y) dy$$

where:

y is the realized county yield

α is the elected coverage level

y^e is the expected area yield

The empirical premium rate (EPR):

This is the current practice

$$EPR = \frac{E[y - \alpha y^{(h)}]}{\alpha y^{(h)}},$$

where:

$y^{(h)}$ is the average historic crop yield

$$E[y - \lambda y^{(h)}] = \sum_{i=1}^{n} (y_i - y^{(h)})/n$$

y_i is the observed crop yield for year i

Rating the crop insurance contract

The fair premium rate (PR):

$$\text{Indemnity} = \max\left\{0, \frac{\alpha y^e - y}{\alpha y^e}\right\}$$

$$\text{PR} = \frac{1}{\alpha y^e} \int_0^{\alpha y^e} (\alpha y^e - y) f(y) dy$$

where:

y is the realized county yield

α is the elected coverage level

y^e is the expected area yield

The empirical premium rate (EPR):

This is the current practice

$$EPR = \frac{E[y - \alpha y^{(h)}]}{\alpha y^{(h)}},$$

where:

$y^{(h)}$ is the average historic crop yield

$$E[y - \lambda y^{(h)}] = \sum_{i=1}^{n} (y_i - y^{(h)})/n$$

y_i is the observed crop yield for year i

Best space-time models

Best hierarchical model

$$\mu_{st} = \rho_1 y_{s,t-1} + \beta_{0_s}^{(M,I)} + \beta_{1_s}^{(C,S)} t + \beta_{2_s}^{(M,I)} t^2 + \mathrm{I}_t^{(M,I)} + \sum_{z \in \{2,4,5,6\}} \beta_{zs}^{(F)} \, WRSI_{zst}$$

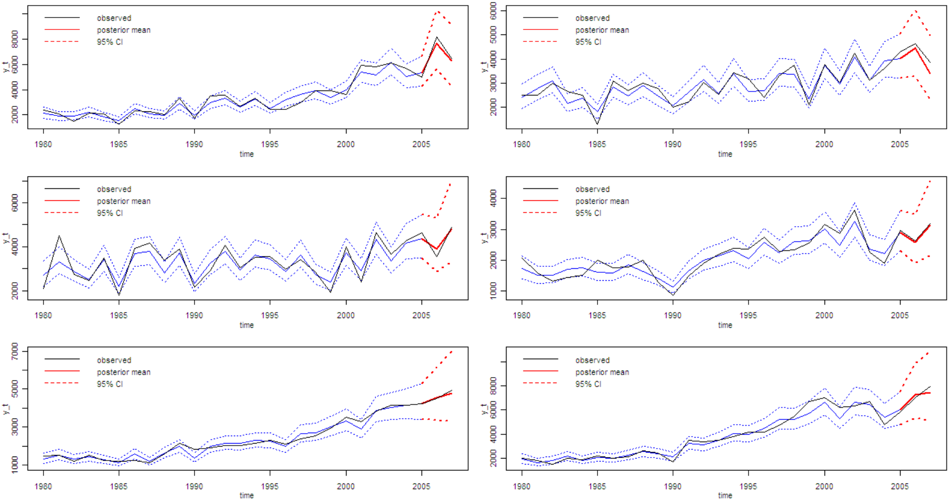
notation: **F**ixed , **C**ounty, **M**icroregion, **I**ID, **S**patial

Preliminary "best" dynamic model

$$\begin{aligned}
\boldsymbol{y}_t &= \boldsymbol{x}_{1t} + \sum_{z \in \{2,4,5,6\}} \Upsilon_{z_t}^{(F)} \, WRSI_{zt} & \boldsymbol{\nu}_t &\sim N\left(\mathbf{0s}, \boldsymbol{V}^{-1}\right) \\
\boldsymbol{x}_{1t} &= \boldsymbol{x}_{1,t-1} + \boldsymbol{x}_{2,t-1} + \boldsymbol{\omega}_{1_t}, & \boldsymbol{\omega}_{1_t} &\sim N\left(\mathbf{0s}, \boldsymbol{W}_1^{-1}\right) \\
\boldsymbol{x}_{2t} &= \boldsymbol{x}_{2,t-1} + \boldsymbol{\omega}_{2_t}, & \boldsymbol{\omega}_{2_t} &\sim N\left(\mathbf{0s}, \boldsymbol{W}_2^{-1}\right)
\end{aligned}$$

Model	ω**MAPE**	ω**MASE**	ω**RMSE**	ω**MAE**
Hierarchical	0.0313	0.2729	47.8757	1.8311
Dynamic	0.0448	0.2462	50.6236	1.9129

Forecast performance in some areas:



Premium rates

Table: Premium rates (%) obtained from the Bayesian hierarchical approach

area	coverage (70%) mean	coverage (70%) MC error*	coverage (90%) mean	coverage (90%) MC error*
Castro	0.014	0.000027	2.041	0.000834
Antonina	7.295	0.001542	16.690	0.002120

* Monte Carlo error of premium rates

- Standard errors of premium rate estimates provide a natural metric to guide the specification of loading factors.

Parallel / Future work

- **Spatio-temporal non-parametric modelling of crop yield through Dirichlet processes;**
- **Zone-based crop insurance;**
- **Seasonal weather forecasts from regional models (ETa) and weather analogue techniques as input in crop simulation models to forecast crop yields;**
- **Remote sensing products (moisture availability indices, etc.) and ENSO indices as additional covariates in crop yield models;**
- **Risk reduction measures (stochastic dominance, VaR, etc.) to evaluate the performance of the proposed approach in reducing risk;**
- **An R package implementing the methodology outlined here is coming soon!**

References

Junger, W.L., Ponce de Leon, A. and Santos, N. (2003) Missing data imputation in multivariate time series via EM algorithm. *Cadernos do IME*, **15**, 8–21.

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

Ruiz-Cárdenas, R., Krainski, E.T. and Rue, H. (2010) Fitting dynamic models using integrated nested Laplace approximations - INLA. NTNU, Preprint Statistics No 12/2010 (submitted).

Sansó, B. and Guenni, L. (2000) A non-stationary multi-site model for rainfall. *Journal of the American Statistical Association*, **95**, 1089–1100.

Simolo, C. Brunetti, M., Maugeria, M. and Nannia, T. (2010) Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. *International Journal of Climatology*, **30**, 1564–1576.

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

Thank you !

visit our websites:

http://www.lce.esalq.usp.br/seguroagricola

http://leg.ufpr.br/projetos:procad