Analysis of Economic Data with Multiscale Spatio-Temporal Models

Marco A. R. Ferreira (University of Missouri - Columbia)

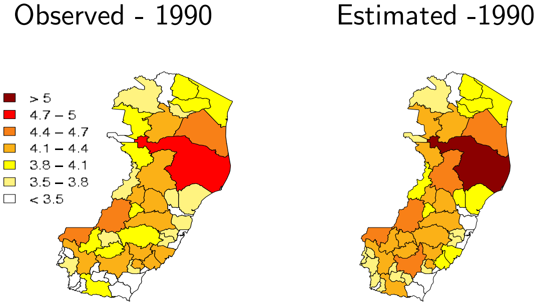
Adelmo Bertolde (Federal University of Espírito Santo, Brazil)

Scott Holan (University of Missouri, Columbia)

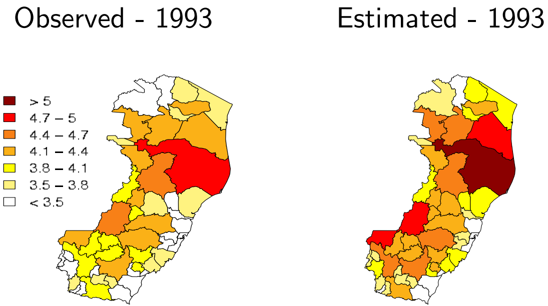
Outline

Outline

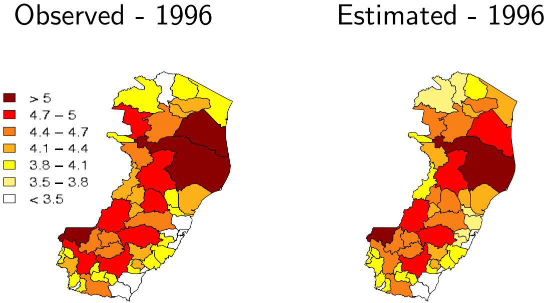
Espírito Santo: Log of agriculture production per county



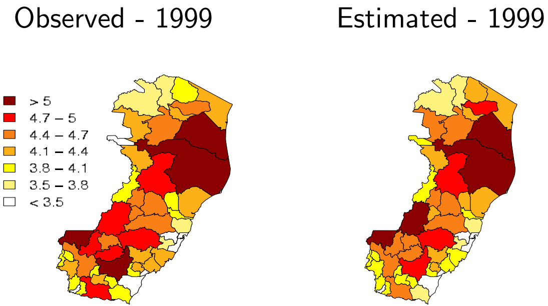
Espírito Santo: Log of agriculture production per county



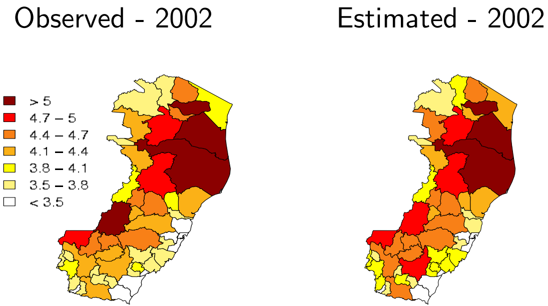
Espírito Santo: Log of agriculture production per county



Espírito Santo: Log of agriculture production per county

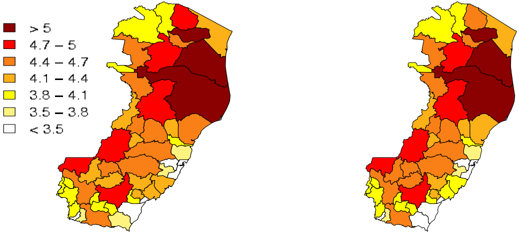


Espírito Santo: Log of agriculture production per county



Espírito Santo: Log of agriculture production per county

Observed - 2005 | Estimated - 2005



Outline

Some background

- Many processes of interest are naturally spatio-temporal.
- Frequently, quantities related to these processes are available as areal data.
- These processes may often be considered at several different levels of spatial resolution.
- Related work on dynamic spatio-temporal multiscale modeling: Berliner, Wikle and Milliff (1999), Johannesson, Cressie and Huang (2007).

Data Structure

Here, the region of interest is divided in geographic subregions or blocks, and the data may be averages or sums over these subregions.

Each state in Brazil is divided into counties, microregions and macroregions; counties are then grouped into microregions, which are then grouped into macroregions, according to their socioeconomic similarity. Thus, our analysis considers three levels of resolution: county, microregion, and macroregion.

Geopolitical organization

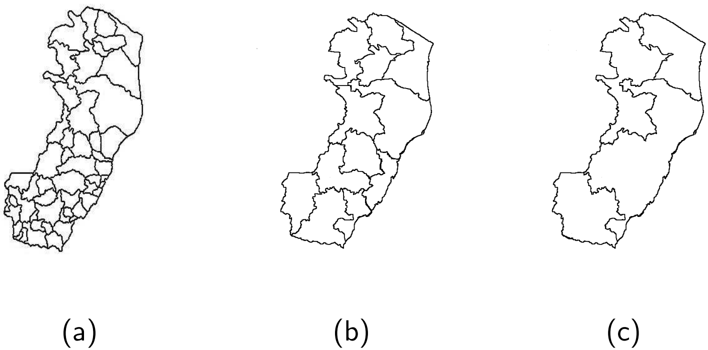


Figure: Geopolitical organization of Espírito Santo State by (a) counties, (b) microregions, and (c) macroregions.

Outline

Multiscale factorization

At each time point we decompose the data into empirical multiscale coefficients using the spatial multiscale modeling framework of Kolaczyk and Huang (2001). See also Chapter 9 of Ferreira and Lee (2007).

Interest lies in agricultural production observed at the county level, which we assume is the L^{th} level of resolution (i.e. the finest level of resolution), on a partition of a domain $S \subset \mathbb{R}^2$.

For the j^{th} county, let y_{Lj}, $\mu_{Lj} = E(y_{Lj})$, and $\sigma^2_{Lj} = V(y_{Lj})$ respectively denote agricultural production, its latent expected value and variance.

Let D_{lj} be the set of descendants of subregion (l, j).
The aggregated measurements at the l^{th} level of resolution are recursively defined by

$$y_{lj} = \sum_{(l+1,j') \in D_{lj}} y_{l+1,j'}.$$

Analogously, the aggregated mean process is defined by

$$\mu_{lj} = \sum_{(l+1,j') \in D_{lj}} \mu_{l+1,j'}.$$

Assuming conditional independence,

$$\sigma^2_{lj} = \sum_{(l+1,j') \in D_{lj}} \sigma^2_{l+1,j'}.$$

Then

$$\mathbf{y}_{D_{lj}} \Big| y_{lj}, \boldsymbol{\mu}_L, \boldsymbol{\sigma}_L^2 \sim N(\boldsymbol{\nu}_{lj} y_{lj} + \boldsymbol{\theta}_{lj}, \boldsymbol{\Omega}_{lj}),$$

with

$$\begin{aligned} \boldsymbol{\nu}_{lj} &= \boldsymbol{\sigma}_{D_{lj}}^2 / \sigma_{lj}^2, \\ \boldsymbol{\theta}_{lj} &= \boldsymbol{\mu}_{D_{lj}} - \boldsymbol{\nu}_{lj} \mu_{lj}, \end{aligned}$$

and

$$\boldsymbol{\Omega}_{lj} = \boldsymbol{\Sigma}_{D_{lj}} - \sigma_{lj}^{-2} \boldsymbol{\sigma}_{D_{lj}}^2 \left(\boldsymbol{\sigma}_{D_{lj}}^2 \right)'.$$

Consider

$$\boldsymbol{\theta}_{lj}^{e} = \mathbf{y}_{D_{lj}} - \boldsymbol{\nu}_{lj} y_{lj},$$

which is an empirical estimate of $\boldsymbol{\theta}_{lj}$.

Then

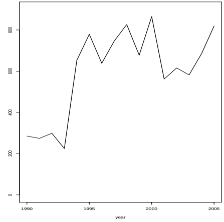
$$\boldsymbol{\theta}_{lj}^{e} | y_{lj}, \boldsymbol{\mu}_L, \boldsymbol{\sigma}_L^2 \sim N(\boldsymbol{\theta}_{lj}, \boldsymbol{\Omega}_{lj}),$$

which is a singular Gaussian distribution (Anderson, 1984).

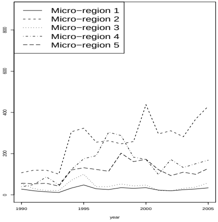
Outline

Exploratory Multiscale Data Analysis

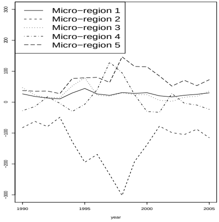
Macroregion 1
total



Disaggregated
by microregion

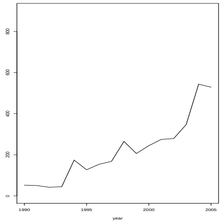


Empirical
multiscale coefficient

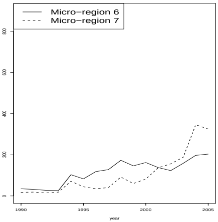


Espírito Santo: Agriculture production of Macroregion 1.

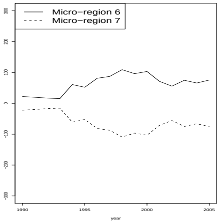
Macroregion 2 total



Disaggregated by microregion

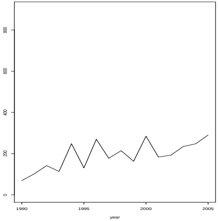


Empirical multiscale coefficient

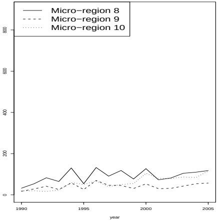


Espírito Santo: Agriculture production of Macroregion 2.

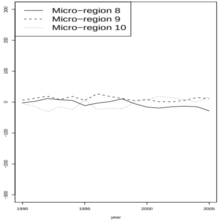
Macroregion 3 total



Disaggregated by microregion

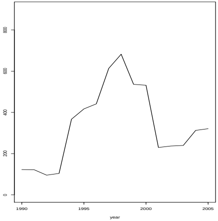


Empirical multiscale coefficient

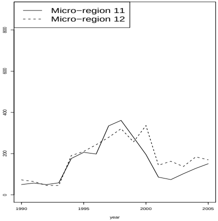


Espírito Santo: Agriculture production of Macroregion 3.

Macroregion 4 total



Disaggregated by microregion



Empirical multiscale coefficient

Espírito Santo: Agriculture production of Macroregion 4.

Outline

The multiscale spatio-temporal model

Observation equation:

$$\mathbf{y}_{tL} = \boldsymbol{\mu}_{tL} + \mathbf{v}_{tL}, \quad \mathbf{v}_{tL} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_L)$$

where

$$\boldsymbol{\Sigma}_L = diag(\sigma^2_{L1}, \ldots, \sigma^2_{Ln_L}).$$

Multiscale decomposition of the observation equation:

$$y_{t1k} \mid \mu_{t1k} \sim N(\mu_{t1k}, \sigma^2_{1k})$$

$$\boldsymbol{\theta}^e_{tlj} \mid \boldsymbol{\theta}_{tlj} \sim N(\boldsymbol{\theta}_{tlj}, \boldsymbol{\Omega}_{lj})$$

System equations:

$$\mu_{t1k} = \mu_{t-1,1k} + w_{t1k}, \quad w_{t1k} \sim N(0, \xi_k \sigma^2_{1k})$$

$$\boldsymbol{\theta}_{tlj} = \boldsymbol{\theta}_{t-1,lj} + \boldsymbol{\omega}_{tlj}, \quad \boldsymbol{\omega}_{tlj} \sim N(\mathbf{0}, \psi_{lj}\boldsymbol{\Omega}_{lj})$$

Priors

$$\mu_{01k}|D_0 \sim N(m_{01k}, c_{01k}\sigma^2_{1k}),$$

$$\boldsymbol{\theta}_{0lj}|D_0 \sim N(\mathbf{m}_{0lj}, C_{0lj}\boldsymbol{\Omega}_{lj}),$$

$$\xi_k \sim IG(0.5\tau_k, 0.5\kappa_k),$$

$$\psi_{lj} \sim IG(0.5\varrho_{lj}, 0.5\varsigma_{lj}),$$

Outline

Empirical Bayes estimation of $\boldsymbol{\nu}_{lj}$ and $\boldsymbol{\Omega}_{lj}$

$\boldsymbol{\nu}_{lj}$: vector of relative volatilities of the descendants of (l, j),
$\boldsymbol{\Omega}_{lj}$: singular covariance matrix of the empirical multiscale coefficient of subregion (l, j)

In order to obtain an initial estimate of σ^2_{Lj}, we perform a univariate time series analysis for each county using first-order dynamic linear models (West and Harrison, 1997). These analyses yield estimates $\tilde{\sigma}^2_{Lj}$.

We estimate $\boldsymbol{\nu}_{lj}$ and $\boldsymbol{\Omega}_{lj}$ by

$$
\begin{aligned}
\tilde{\boldsymbol{\nu}}_{lj} &= \tilde{\boldsymbol{\sigma}}^2_{D_{lj}} / \tilde{\sigma}^2_{lj}, \\
\tilde{\boldsymbol{\Omega}}_{lj} &= \tilde{\boldsymbol{\Sigma}}_{D_{lj}} - \tilde{\sigma}_{lj}^{-2}.
\end{aligned}
$$

Outline

Posterior exploration

Let

$$
\begin{aligned}
\boldsymbol{\theta}_{\bullet lj} &= (\boldsymbol{\theta}'_{0lj}, \ldots, \boldsymbol{\theta}'_{Tlj})', \\
\boldsymbol{\theta}_{t\bullet j} &= (\boldsymbol{\theta}'_{t1j}, \ldots, \boldsymbol{\theta}'_{tLj})', \\
\boldsymbol{\theta}_{\bullet\bullet\bullet} &= (\boldsymbol{\theta}'_{\bullet 11}, \ldots, \boldsymbol{\theta}'_{\bullet 1n_1}, \boldsymbol{\theta}'_{\bullet 21}, \ldots, \boldsymbol{\theta}'_{\bullet 2n_2}, \ldots, \boldsymbol{\theta}'_{\bullet L1}, \ldots, \boldsymbol{\theta}'_{\bullet Ln_L})',
\end{aligned}
$$

with analogous definitions for the other quantities in the model.

It can be shown that, given $\boldsymbol{\sigma}^2_\bullet$, $\xi_\bullet$, and $\psi_{\bullet\bullet}$, the vectors $\boldsymbol{\mu}_{\bullet 11}, \ldots, \boldsymbol{\mu}_{\bullet 1n_1}$, $\boldsymbol{\theta}_{\bullet 11}, \ldots, \boldsymbol{\theta}_{\bullet 1n_1}, \ldots, \boldsymbol{\theta}_{\bullet L1}, \ldots, \boldsymbol{\theta}_{\bullet Ln_L}$, are conditionally independent *a posteriori*.

Gibbs sampler

- $\boldsymbol{\mu}_{\bullet 1k}$: Forward Filter Backward Sampler (FFBS) (Carter and Kohn, 1994; Fruhwirth-Schnatter,1994).
- $\xi_k | \mu_{\bullet 1k}, \sigma^2_{1k}, D_T \sim IG\left(0.5\tau^*_k, 0.5\kappa^*_k\right)$, where $\tau^*_k = \tau_k + T$ and $\kappa^*_k = \kappa_k + \sigma^{-2}_{1k} \sum_{t=1}^{T} (\mu_{t1k} - \mu_{t-1,1k})^2$.
- $\psi_{lj} | \boldsymbol{\theta}_{\bullet lj}, D_T \sim IG(0.5\varrho^*_{lj}, 0.5\varsigma^*_{lj})$, where $\varrho^*_{lj} = \varrho_{lj} + T(m_{lj} - 1)$ and $\varsigma^*_{lj} = \varsigma_{lj} + \sum_{t=1}^{T} (\boldsymbol{\theta}_{tlj} - \boldsymbol{\theta}_{t-1,lj})' \boldsymbol{\Omega}^-_{lj} (\boldsymbol{\theta}_{tlj} - \boldsymbol{\theta}_{t-1,lj})$, where $\boldsymbol{\Omega}^-_{lj}$ is a generalized inverse of $\boldsymbol{\Omega}_{lj}$.
- $\boldsymbol{\theta}_{\bullet lj}$: Singular FFBS.

Singular FFBS

1. Use the Kalman filter to obtain the mean and covariance matrix of $f(\boldsymbol{\theta}_{1lj}|\sigma^2, \psi_{lj}, D_1), \ldots, f(\boldsymbol{\theta}_{Tlj}|\sigma^2, \psi_{lj}, D_T)$:
   - posterior at $t-1$: $\boldsymbol{\theta}_{t-1,lj}|D_{t-1} \sim N\left(\mathbf{m}_{t-1,lj}, C_{t-1,lj}\boldsymbol{\Omega}_{lj}\right)$;
   - prior at t: $\boldsymbol{\theta}_{tlj}|D_{t-1} \sim N\left(\mathbf{a}_{tlj}, R_{tlj}\boldsymbol{\Omega}_{lj}\right)$, where $\mathbf{a}_{tlj} = \mathbf{m}_{t-1,lj}$ and $R_{tlj} = C_{t-1,lj} + \psi_{lj}$;
   - posterior at t: $\boldsymbol{\theta}_{tlj}|D_t \sim N\left(\mathbf{m}_{tlj}, C_{tlj}\boldsymbol{\Omega}_{lj}\right)$, where $C_{tlj} = (1 + R_{tlj}^{-1})^{-1}$ and $\mathbf{m}_{tlj} = C_{tlj}\left(\boldsymbol{\theta}^e_{tlj} + R_{tlj}^{-1}\mathbf{a}_{tlj}\right)$.
2. Simulate $\boldsymbol{\theta}_{Tlj}$ from $\boldsymbol{\theta}_{Tlj}|\sigma^2, \psi_{lj}, D_T \sim N(\mathbf{m}_{Tlj}, C_{Tlj}\boldsymbol{\Omega}_{lj})$.
3. Recursively simulate $\boldsymbol{\theta}_{tlj}$, $t = T-1, \ldots, 0$, from

$$\boldsymbol{\theta}_{tlj}|\boldsymbol{\theta}_{t+1,lj}, \ldots, \boldsymbol{\theta}_{Tlj}, D_T \;\equiv\; \boldsymbol{\theta}_{tlj}|\boldsymbol{\theta}_{t+1,lj}, D_t \;\sim\; N(\mathbf{h}_{tlj}, H_{tlj}\boldsymbol{\Omega}_{lj}),$$

where $H_{tlj} = \left(C_{tlj}^{-1} + \psi_{lj}^{-1}\right)^{-1}$ and

$$\mathbf{h}_{tlj} = H_{tlj}\left(C_{tlj}^{-1}\mathbf{m}_{tlj} + \psi_{lj}^{-1}\boldsymbol{\theta}_{t+1,lj}\right).$$

Reconstruction of the latent mean process

One of the main interests of any multiscale analysis is the estimation of the latent mean process at each scale of resolution.

From the g^{th} draw from the posterior distribution, we can recursively compute the corresponding latent mean process at each level of resolution using the equation

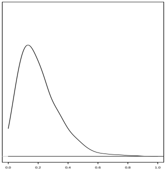
$$\boldsymbol{\mu}_{t,D_{lj}}^{(g)} = \boldsymbol{\theta}_{tlj}^{(g)} + \boldsymbol{\nu}_{tlj}\mu_{tlj}^{(g)},$$

proceeding from the coarsest to the finest resolution level.

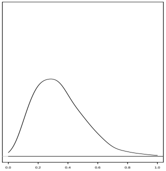
With these draws, we can then compute the posterior mean, standard deviation and credible intervals for the latent mean process.

Outline

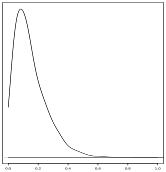
Marginal posterior densities for the signal-to-noise ratio ξ_k



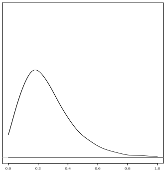
ξ_1



ξ_2

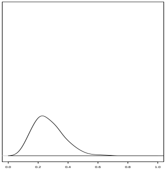


ξ_3

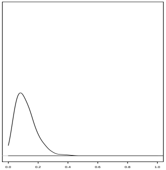


ξ_4

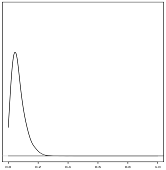
Marginal posterior densities for the signal-to-noise ratio ψ_{1j}



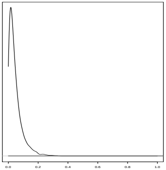
ψ_{11}



ψ_{12}

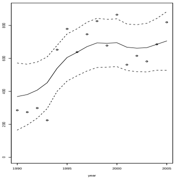


ψ_{13}

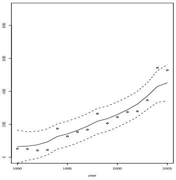


ψ_{14}

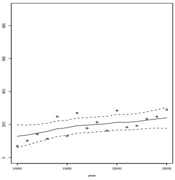
Mean process at coarse level



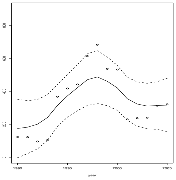
μ_{t11}



μ_{t12}

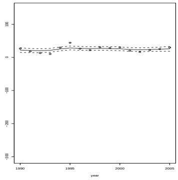


μ_{t13}

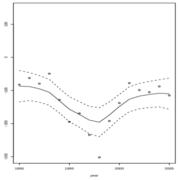


μ_{t14}

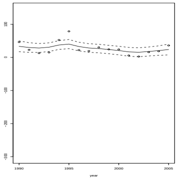
Multiscale coefficient for Macroregion 1



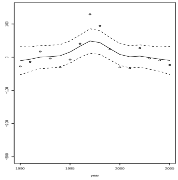
θ_{t111}



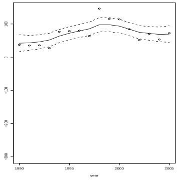
θ_{t112}



θ_{t113}

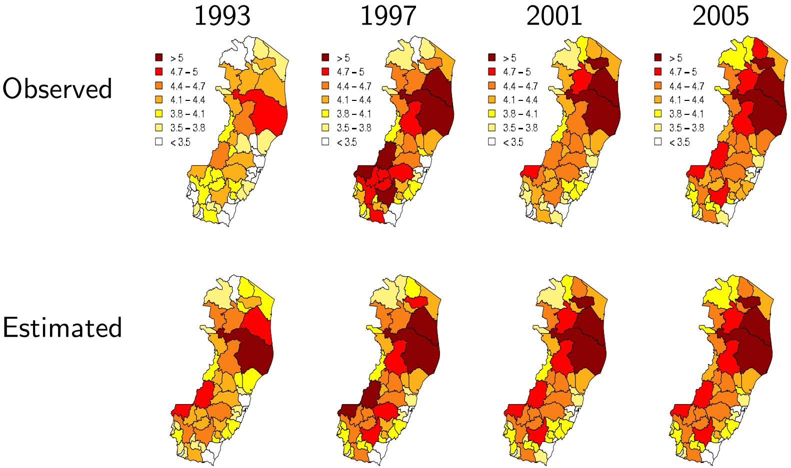


θ_{t114}



θ_{t115}

Observed agriculture production and estimated mean



Outline

Conclusions

- New multiscale spatio-temporal model for areal data.
- Dynamic multiscale coefficients.
- Efficient Bayesian estimation.
- Potential to be used with massive datasets.