A State-Space Model for Understanding Spatial Dynamics Represented by Areal Data

Guofeng Cao\textsuperscript{1}, Shaowen Wang\textsuperscript{1,2,3,4}, Qingfeng Guan\textsuperscript{5}

\textsuperscript{1}Department of Geography
\textsuperscript{2}Department of Computer Science
\textsuperscript{3}Department of Urban and Regional Planning
\textsuperscript{4}National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign, 61801
Email: \{guofeng; shaowen\}@illinois.edu

\textsuperscript{5}Center for Advanced Land Management Information Technologies
School of Natural Resources
University of Nebraska - Lincoln
Email: qguan2@unl.edu

1. Introduction

Most geographical phenomena, if not all of them, are essentially dynamic processes instead of static forms. The continuing advances of data acquisition and dissemination technologies have made available massive amounts of spatiotemporal data for the research of such dynamic processes. In the past decade, a surge of statistical models have been proposed in GIScience and related fields for modeling geo-referenced spatiotemporal data, with point support in particular, by accounting for complex spatiotemporal interactions. Much less attention, however, has been given for the spatiotemporal data collected and aggregated over geographical units (areal data), which are all-important and common data sources in many scientific disciplines. Many existing methods are developed particularly for spatiotemporal disease mapping (e.g., Schmid & Held 2004), and they tend to assume that areal units share a common and well-behaved temporal trend. These assumptions hinder the applications of such methods in dynamics with complex and heterogeneous spatiotemporal interactions, which (unfortunately) are typical in GIScience.

State-space models provide a unifying and intuitive framework for analyzing behaviors of dynamic systems. Such models allow researchers to model the dynamics in a time series by a sequence of (often latent) state variables driven by a stochastic process. Gaussian Markov Random Field (GMRF) models, also known as conditional autoregressive models, are arguably the most commonly used methods for spatial modeling of areal data. GMRF models basically assume that a variable associated with an areal unit primarily depends on its neighborhood (Markov property), and the spatial dependence structure is described via a precision matrix. Inspired by Vivar & Ferreira (2009), this paper represents each entry of an areal data series as a GMRF and integrates them into a compact state-space form to investigate the spatial dynamics of areal data. Specifically, a latent random field (GMRF distributed) process is assumed to evolve under the time series of areal data in a fashion of autoregressive process, and at each time instant, this latent random field is connected to the observable areal variables via a link function with spatially correlated (again GMRF distributed) errors. Both complex spatial interactions (via GMRFs) and temporal interactions (via an autoregressive process in state-spaces modes) are thus integrated in this model. For model inferences, a recently proposed Integrated Nested Laplace Approximation (INLA) (Rue et al. 2009), instead of Markov Chain Monte Carlo (MCMC) as in Vivar & Ferreira (2009), is adopted to avoid the computational and convergence issues involved in the use of MCMC.
2. Methodology

Consider a geographical area, which is divided into $N$ disjoint areal units with defined boundaries. Topologically, such a division can be represented as a connected graph with each unit $i$ having a neighbourhood $\partial_i$ defined by contiguities. Over a number of time instants $t_1, \ldots, t_T$, we repeatedly sample on a variable $Y$, and denote the outcome as $y_i = \{y_{it}: i = 1, \ldots, N, t = t_1, \ldots, t_T\}$. Along with $Y$, we assume that there is a latent random field process $z_1, \ldots, z_T$ in the same area, and that given this process, $y_i$ are independent with each other and link to $z_i$ through Equation (1.1), namely the observation equation in state-space models.

$$y_i = Hz_i + v_i \quad v_i \sim \mathcal{N}(0, \Sigma_v^{-1}(\theta_v))$$

In this equation, $H$ represents the correlation matrix between the latent random field process and $y_i$, and the error $v_i$ is a GMRF specified by a precision matrix $\Sigma_v^{-1}$ controlled by a hyper-parameter $\theta_v$, which is assumed to be fixed over the evolution of the process. If let $H$ be the identity matrix $I$, $z_i$ can then be taken as the spatial trend field of observations at time $t$.

To model temporal interactions, we assume latent fields $z_1, \ldots, z_T$ follow an autoregressive process defined by Equation (1.2), or the state equation in state-space models.

$$z_i = Bz_{i-1} + w_i \quad w_i \sim \mathcal{N}(0, \Sigma_w^{-1}(\theta_w))$$

The evolution matrix $B$ specifies the complexity (order) and characteristic of temporal interactions; same as $v_i$, $w_i$ is also a GMRF distributed error term, whose characteristics are specified by a precision matrix $\Sigma_w^{-1}$, and $\theta_w$ is also assumed to be fixed. The first-order autoregressive process has $B = \rho I$. Further let coefficient $\rho = 1$, we have a random walk process for each unit. Figure 1 gives a simple illustration of the evolution process of this model.

![Figure 1: An illustration of a spate-space model.](image)

Complexity and behavior of this model are contingent on the specification of parameters $\{H, B, \theta_v, \theta_w\}$, where $H$ and $B$ are often assumed to be known by modellers.
To infer the latent process $z_1, \ldots, z_T$ and the unknown hyper-parameter $\theta_v$ and $\theta_w$, Vivar & Ferreira (2009) proposed a method based on Markov chain Monte Carlo (MCMC) framework. MCMC-based methods, however, often suffer from their issues such as convergence and heavy computational load. Following Ruiz-Cárdenas et al. (2011), this paper adopts a fast alternative to MCMC, so-called Integrated Nested Laplace Approximation (INLA), for model inference.

3. Case Study

Drought has been a serious natural disaster in the United States. The Risk Management Agency (RMA) of the United States Department of Agriculture (USDA) has been collecting the indemnity payments for agricultural losses due to various reasons including drought (http://www.rma.usda.gov/data/cause.html). As an example to showcase the utility of the proposed model, spatiotemporal trends of drought-caused indemnity payment in the counties of Texas in the past 11 years (2001~2011) were analysed. In this case study, logarithm values of the payment data (for positive numbers) were used (a snap shot of year 2001 is shown in Figure 2). We adopted a random walk process in Equation (1.2), and a first-order of contiguity neighborhood in a GMRF. We further let $H = I$, and as mentioned in the previous section, the latent random field $z_t$ thus represents the spatial trend field at time $t$. The trend of spatial dynamics can thus be interrogated through the trajectories of $z_t$. The analysis was performed using the INLA package in the R statistical computing environment.

![Figure 2: Drought Indemnity Map of Texas, year 2001](image)

Figure 3 shows the evolution process of the indemnity values in Bosque County and its seven neighboring counties (Coryell, Erath, Hamilton, Hill, Johnson, McLennan and Somervell). We can clearly see that the overall temporal trends ($z_{it}$) are well captured in each of these counties. Figure 4 shows the indemnity values (left column) and the estimated spatial trends (right column) of year 2003~2007. Evidently, the spatial trends well reflect the indemnity distribution. Over the years, the northwest and southeast counties of Texas suffered higher indemnity than other counties. In year 2005, however, the indemnity values of southeast areas (indicated by the red circles in Figure 4) are dramatically lower than those of other years. The northwest areas performed in the same fashion in year 2006. These dramatic changes (which
can be considered as outliers) did not significantly influence the spatial trend of those two years. It makes senses actually considering the smoothing effect of temporal trends, as demonstrated in Figure 3, and this effect makes the model tolerant of noises, errors and outliers in observations.

Figure 3: Temporal trends of indemnity values in neighboring counties. Red solid lines are for temporal trends, black solid lines for observed indemnity values, and the blue dash lines for 95% confidence intervals of the estimated temporal trends.
Figure 4: Indemnity maps (left column) and estimated spatial trends (right column) of year 2003–2007. In year 2005, the indemnity values of southeast counties, located in the red circle, are dramatically lower than the rest years. Same for the northwest counties in year 2006.
To further demonstrate the prediction performance of the model, we assume the values of several counties are missing for years 2006~2011, and use the previous years (2001~2005) of data of these counties as well as neighboring counties data to predict them. As Figure 5 shows, the missing values (at Austin, Briscoe, Dawson and Fisher) are well recovered.

![Figure 5: Predictions of missing values. Red solid lines are for prediction values, black solid lines for observed indemnity values, and the blue dash lines for 95% confidence intervals of the predictions.](image)

4. Discussion

It is essential to understand the spatiotemporal characteristics of areal data in GIScience. This paper described a model to investigate the complex spatial-temporal interactions in areal data by integrating a Gaussian Markov Random Field model (for spatial interactions) into a dynamic state-spaces model (for temporal interactions). A case study was conducted to showcase the advantages of the proposed model, particularly in characterizing spatio-temporal trends and estimating uninformed areal units. In practice, time-dependent covariates are often available, and a successful integration of such auxiliary information is anticipated to further improve the understanding of spatial dynamics represented in areal data, which will be investigated in the future work based on the current findings.

Acknowledgements

Funding provided for this work by the National Science Foundation (NSF) under award SI2-SSI #1047916 is gratefully acknowledged.

References

