5th International Workshop on Climate Informatics
September 24-25, 2015
Hosted by the National Center for Atmospheric Research in Boulder, CO

# SPATIO-TEMPORAL HYPOTHESIS TESTING IN MODEL RESIDUALS

Lindsey R. Dietz[1], Snigdhansu Chatterjee[1]

***Abstract*—Spatio-temporal models add complexity, but not necessarily value, to some climate analyses. To confirm the presence of spatio-temporal dependence, a hypothesis test should be conducted. The *Space-Time Index* is one statistic to detect such dependence; this statistic is simple, easily interpretable, and used in several disciplines. In an application to Indian monsoon precipitation thresholds, residuals from logit-normal mixed models were tested for spatio-temporal dependence. No evidence of dependence was detected in high thresholds.**

## I. SPACE-TIME INDEX (STI) METHODOLOGY

Spatio-temporal dependence in climate related data should not be ignored in modeling efforts. Along with graphical diagnostics, it is advantageous to have formal hypothesis testing procedures in place to understand the exact nature of spatio-temporal dependence, and to evaluate whether a given statistical model is adequate in capturing such dependencies in the data. Elegant procedures exist for testing separability, symmetry [1], or stationarity [2] in the data.

Another simple method still in current use ([3], [4], [5]) is the Space-Time Index (STI) [6] which combines Moran's I [7] and the Durbin-Watson statistic [8]. The STI is interpretable and is useful for conveying information to stakeholders who may not be experts on spatio-temporal data patterns and related statistical models. Our goal is to evaluate the performance of STI for modeling dependencies in climate data. Based on that, we will propose and conduct future studies on computational methodology-based generalizations to overcome the limitations of STI as it applies to non-stationary, high-dimensional data from climate applications.

In its current form, the STI tests the null hypothesis of no spatio-temporal dependence in a vector autoregression process over the entire spatial field given by: $\mathbf{y}_t = A\mathbf{y}_{t-1} + \epsilon_t$ where $t \in \{1, ..., T\}$ represents discrete time. Let $i, j \in \{1, ..., n\}$ represent stations, $\bar{y} = \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} y_{i,t}$, and $c_{ij,t-1} = 1$ if stations $i$ and $j$ are neighbors during time $t-1$ and $c_{ij,t-1} = 0$ otherwise. Then,

Corresponding author: L.R. Dietz (diet0146@umn.edu), School of Statistics, University of Minnesota, Minneapolis, MN[1]

$$STI = \frac{n(T-1)}{\sum_{t=2}^{T}\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij,t-1}} * \frac{\sum_{t=2}^{T}\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij,t-1}(y_{i,t}-\bar{y})(y_{j,t-1}-\bar{y})}{\sum_{t=1}^{T}\sum_{i=1}^{n}(y_{i,t}-\bar{y})^2}$$

Under an asymptotic normality assumption, the sampling distribution of STI can be used to conduct the test.

## II. STI SIMULATION RESULTS

Structural assumptions are imposed to run simulations. First, assume the neighbors of a point remain constant, i.e. $c_{ij,t-1} = c_{ij,t}$ for all $t$. Neighbors of station $i$ ($N_i$) are weighted by scaled distances ($\mathbf{w}_i$) where $\mathbf{w}_i$ satisfies:

1) $w_{ij}$=0 if j $\notin N_i$, $w_{ij}$=$\sum_{j\in N_i} \frac{dist_j}{\sum_{j\in N_i} dist_j}$ for $j \in N_i$
2) $\sum_{j=1}^{n} w_{ij} = 1$ for all $i$.

Next, generate the $i^{th}$ vector with time parameter ($\rho_{time}$) and space parameter ($\rho_{space}$) as:

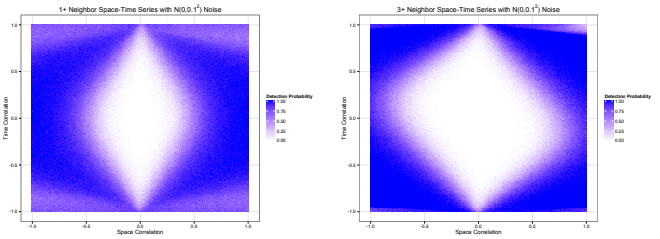
$$t = 1 : y_{i,1} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2) \text{ for all } i$$

$$t > 1 : y_{i,t} = \rho_{time} \cdot y_{i,t-1} + \rho_{space} \cdot \sum_{t=2}^{T}\sum_{j\in N_i}(w_{ij} \cdot y_{j,t-1}) + \epsilon_{i,t}$$

$$\epsilon_{i,t} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2) \text{ for all } i, t$$

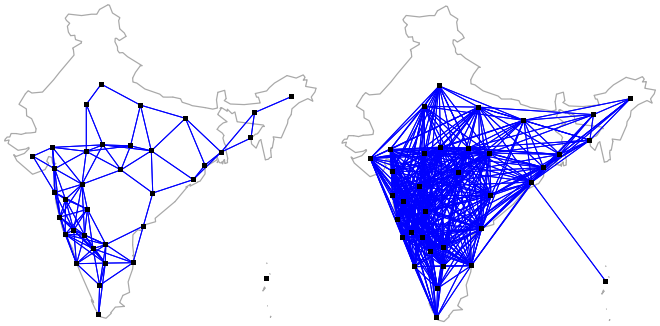
In one setting, a 100 point time series for a 3x3 grid of spatial locations was generated independently 100 times with different neighbor networks. Detection of spatio-temporal dependence was defined as obtaining a p-value $< 0.05$. As seen in Fig. 1, the power of the test was low, especially as the number of spatial neighbors increased. There was also a failure to detect some of the highest combinations of correlations as seen in the corners of the figures.

Fig. 1. $\geq 1$ and $\geq 3$ Neighbor Detection of Space-Time Correlation



5th International Workshop on Climate Informatics
September 24-25, 2015
Hosted by the National Center for Atmospheric Research in Boulder, CO

Fig. 2. Neighbors Network for Stations within 500KM & 1250KM



## III. STI Application in Indian Monsoon Precipitation Modeling

We apply STI to the logit-normal model residuals produced in [9]. Data specifications remain the same, using daily data for monsoon seasons from 1973-2013.

Let station $i \in \{1, ..., m\}$, day $d \in \{1, ..., n_i\}$, and year $k \in \{1973, ..., 2013\}$. Given a precipitation (mm·day$^{-1}$) threshold $\tau \in \{50, 75, 100, 125\}$ and daily precipitation event $Z_{idk}$, let $T_{idk} = I(Z_{idk} > \tau)$. Let $\mathbf{x}_{ijk}$ be a vector of covariates and $\mathbf{U}$ and $\mathbf{W}$ be vectors of random effects for station and year, respectively. Then, the logit-normal mixed model is:

$$\begin{aligned}
&\text{Level 1: } T_{idk}|\mathbf{U} = \mathbf{u}, \mathbf{W} = \mathbf{w} \overset{\text{ind.}}{\sim} \text{Bernoulli}(\theta_{ijk}),\\
&\qquad \text{logit}(\theta_{idk}) = \mathbf{x}_{idk}^{\mathrm{T}}\boldsymbol{\beta} + u_i + w_k,\\
&\text{Level 2: } U_i \overset{\text{ind.}}{\sim} \mathcal{N}(0, \sigma^2_{station}), W_k \overset{\text{ind.}}{\sim} \mathcal{N}(0, \sigma^2_{year})\\
&\qquad U_i \text{ independent of } W_k \text{ for all } (i, k).
\end{aligned}$$

The model explicitly accounts for time dependence but only implicitly for spatial dependence, thus, we want to assess the need for further structure.

We collect Pearson-like residuals ($r_{idk}$) for each observation during model estimation within SAS/STAT® 9.3. Because of the model structure, linearized pseudo-data ($\tilde{p}_{idk}$) is computed for comparison to logit($\hat{\theta}_{idk}$) for each observation during model fitting. eBLUPs are used as estimates for random effects $\mathbf{u}$ and $\mathbf{w}$. Thus, the model residuals are:

$$r_{idk} = \frac{\tilde{p}_{idk} - (\mathbf{x}_{idk}^{\mathrm{T}}\hat{\boldsymbol{\beta}} + \widehat{u_i} + \widehat{w_k})}{\widehat{\text{Var}}(\tilde{p}_{idk}|\mathbf{u}, \mathbf{w})}.$$

Five distance settings are employed resulting in different neighborhoods. Networks for 500KM and 1250KM are seen in Fig. 2.

Table I displays the higher rainfall thresholds do not show evidence of spatio-temporal dependence in the model residuals. Results are clearly influenced by the choices of neighbors. However, no spatio-temporal dependence was detected at any distance for higher rainfall amounts. Taking into account the power of the test, we conclude that if dependence exists, it is not extremely strong and may not require additional structure.

TABLE I
P-values for STI Based on Model Residuals

| | # KM in which Neighbors Exist | | | | |
|---|---|---|---|---|---|
| Rain in mm·day$^{-1}$ | 250 | 500 | 750 | 1000 | 1250 |
| $\geq 50$ | **0.00** | **0.01** | 0.11 | 0.25 | 0.33 |
| $\geq 75$ | **0.00** | 0.16 | 0.59 | 0.67 | 0.61 |
| $\geq 100$ | 0.27 | 0.58 | 0.84 | 0.90 | 0.67 |
| $\geq 125$ | 0.70 | 0.84 | 0.93 | 0.97 | 0.71 |

## IV. Future Work

Although STI provides a useful first effort in identifying spatio-temporal dependence in residuals, testing is currently restrictive in scope. Future research includes:

1) Modification of the hypothesis to test for temporal or spatial correlation separately.
2) Modification of the space-time process to include AR($p$), $p > 1$ correlations.
3) Conducting a permutation test rather than relying on asymptotic normality
4) Investigating cancellation of correlation when spatial and temporal signals have opposite signs

## Acknowledgments

Ms. Dietz is supported by NSF Expeditions in Computing Award #1029711 and a UMN Doctoral Dissertation Fellowship.

## References

[1] B. Li, M. G. Genton, and M. Sherman, "A Nonparametric Assessment of Properties of Space–Time Covariance Functions," *Journal of the American Statistical Association*, vol. 102, pp. 736–744, June 2007.

[2] M. Fuentes, "Testing for separability of spatial–temporal covariance functions," *Journal of Statistical Planning and Inference*, vol. 136, no. 2, pp. 447–466, 2006.

[3] N. Cressie and C. K. Wikle, *Statistics for Spatio-Temporal Data*. Wiley Series in Probability and Statistics, Hoboken, New Jersey: John Wiley and Sons, Inc., 1 ed., 2011.

[4] M.-P. Kwan, D. Richardson, D. Wang, and C. Zhou, eds., *Space-Time Integration in Geography and GIScience*. Netherlands: Springer, 2015.

[5] T. Mitze, *Empirical Modelling in Regional Science*, vol. 657 of *Lecture Notes in Economics and Mathematical Systems*. Berlin Heidelberg: Springer-Verlag Berlin Heidelberg Springer-Verlag, 2012.

[6] D. A. Griffith, "Interdependence in Space and Time : Numerical and Interpretative Considerations," in *Dynamic Spatial Models* (D. A. Griffith and R. MacKinnon, eds.), pp. 258–287, Sijthoff & Noordhoff, 1981.

[7] P. A. P. Moran, "Notes on Continuous Stochastic Phenomena," *Biometrika*, vol. 37, pp. 17–23, June 1950.

[8] J. Durbin and G. S. Watson, "Testing for Serial Correlation in Least Squares Regression: I," *Biometrika*, vol. 37, pp. 409–428, December 1950.

[9] L. R. Dietz and S. Chatterjee, "Investigation of Precipitation Thresholds in the Indian Monsoon Using Logit-Normal Mixed Models," in *Machine Learning and Data Mining Approaches to Climate Science* (V. Lakshmanan, E. Gilleland, A. McGovern, and M. Tingley, eds.), (Cham), pp. 239–246, Springer International Publishing, July 2015.