# Directional Spatial Dependence and Its Implications for Modeling Systemic Yield Risk

Ying Zhu

Department of Agricultural and Resource Economics

North Carolina State University

yzhu@ncsu.edu

Sujit K. Ghosh

Department of Statistics

North Carolina State University

Barry K. Goodwin

Department of Agricultural and Resource Economics

North Carolina State University

*Selected Paper prepared for presentation at the Agricultural & Applied Economics Association 2009 AAEA & ACCI Joint Annual Meeting, Milwaukee, Wisconsin, July 26-29, 2009*

# Directional Spatial Dependence and Its Implications for Modeling Systemic Yield Risk

**Ying Zhu, Sujit K. Ghosh and Barry K. Goodwin**

**Abstract:**

The objective of this study is to evaluate and model the spatial dependence of systemic yield risk. Various spatial autoregressive models are explored to account for county level dependence of crop yields. The results show that the time trend parameters of yields are correlated across spaces and the spatial correlations are changing with time. In addition, the spatial correlation of neighborhood in west/east direction is stronger than that of north/south direction. The information of the spatial dependence of yield risk will help the construction of better risk management programs for protecting producers from systemic yield risks.

**Key Words:** Spatial Autoregressive Model, Spatial Dependence

# 1 Introduction

Yield risk for a given crop can differ systematically over space due to the changing agronomic conditions, such as climate and soil type. The concept of spatial correlation is important in crop insurance due to systemic risk. It's been argued that the causes of crop insurance market failure is systemic risk (Miranda and Glauber (1997)). Systemic risk in agriculture come from the fact that adverse weather events usually induce substantial correlations among counties. As a result, yields are usually spatially correlated and yield shortfalls in a particular area such as a county or state are likely to be correlated with yield shortfalls in the neighboring counties.

U.S. crop insurers face portfolio risk due to high level of systemic risk. High level of systemic risk undermines a crop insurer's ability to diversify risk across space and prevents it from performing the function of pooling risk across farmers. Economic motivation of using spatial models comes from the fact that counties differ in their agricultural production which might be rooted in weather and soil type differences that vary smoothly over space. Modeling the spatial correlation between yield losses in different counties and measurement of the systemic crop yield risk will give implication for insurablity, risk management, reinsurance and crop insurance pricing, as well as help conduct better risk management of systemic risk and maintain consistency of premium rates among contiguous counties. As an example, in the Group Risk Plan (GRP), coverage and indemnities are based on a given geographical area. All farmers in this region pay the same premium and receive the same indemnity per unit of insurance according to a county index that is the basis for determining a loss. A representative area-yield insurance plan requires carefully investigated spatial correlation of crop yield losses so that the reasonable range and shape of the geographical area can be considered in the design of the optimal GRP. Premium rates based on arbitrary smoothing area yields could lead to increased program losses due to adverse selection. Accurately modeling contiguous county information in the construction of insurance contract is thus very important for risk management of systemic crop risk.

The primary goal of this study is to measure the systemic crop loss risk by exploring the spatial dependence structure of crop yields. The focus is on modeling the spatial correlation parameter between crop yields across space (both in contiguity condition and direction). To model the spatial correlation of yields and capture the neighborhood-level spatial effects, an autoregressive model (AR) holds promise. AR model includes the conditional autoregressive model (CAR) and the simultaneously autoregressive model (SAR). An AR model is the model incorporating the discrete neighborhood information and pools innovations from contiguous counties. Thus, it reduces the sampling variance of the estimated conditional distribution of the central county. In the AR model, a neighborhood structure based on the shape of the lattice is defined. Thus, instead of measuring the distance between centroids of regions, a system is used to define regions to be neighbors based on whether their borders touch or not. Once the neighborhood structure is defined, models similar to time series autoregressive models are considered. The estimated rate of spatial correlation may depend not only on the way the regions are connected, but also on the direction that they are separated. This study uses a directional autoregressive (DAR) model to estimate the spatial pattern and the correlation parameter between crop yields through space. Estimation results will have implications on how strong the spatial dependence is for observations from nearby regions to be more (or less) alike than observations from regions farther apart. This information of the variation of the spatial correlation across space can provide insights into the design of the area-yield contract and the agricultural risk management policy.

# 2 The Econometric Model

## 2.1 Spatial Correlation for Lattice Data

In many settings, average yields over geographically defined region such as a county or a state are observed and regression or classification analysis is performed. In general, given a lattice which consist of a set of sub-region $S_1, \ldots, S_I$, a generalized linear model can be used for the aggregated crop yields $Y_t$.

$$Y_t = \mu_t + \eta_t \quad (1)$$

where $Y_t = (y_{1t}, \ldots, y_{It}) = (y_t(S_1), \ldots, y_t(S_I))$ are area yields and $\mu_t = (\mu_{1t}, \ldots, \mu_{It})$ represents a vector of large scale variations (for example, time trend) over region $S_i$. Usually, it is modeled as a deterministic function of some explanatory variables such as time, weather and other area level covariates. In other words, $\mu_{it} = X_i\beta_i$, where $X_i = (X_{i1}, \ldots, X_{iT})$, and $X_i$ are explanatory variables in sub-region $S(i)$; $\beta_i$ is a set of finite dimensional parameter. $\eta_t = (\eta_{1t}, \ldots, \eta_{It})$ represents a vector of small-scale variations (i.e. spatial random effects) with zero mean and variance-covariance $\Sigma_t$ ($\eta_t \sim (0, \Sigma_t)$). In crop insurance literature, the part of modeling the large scale variation is well illustrated. For example, $\mu_t$ can be modeled using parametric or semiparametric model (Goodwin and Ker (2001)). However, it is less discussed in the literature how to model for the spatial random effect $\eta_t$ as they are spatially correlated and the positive definiteness condition of the induced covariance structure needs to be specified. Issues related to spatially autocorrelated disturbance terms need to be considered in estimating the area yields. For example, whether there is any apparent tendency in the residual term that may not come from a random chance alone. The CAR and SAR models are spatial models which include the spatial correlation parameters to control the strength of spatial correlation. These models can be used in modeling the county-level regional yield data.

## 2.2 Conditional and Simultaneously Spatial Autoregressive Model

AR model is a method to model $\mathrm{y}_t$ based on the first-order auto-regression on the average of its neighbors' response. In the AR model, the spatial dependence $\rho$ is incorporated into the covariance structure via an autoregressive model. We could adopt a AR for the deviations of yields from their site–specific means and carry out maximum likelihood to estimate the model.

Let $Y_t = (y_{1t}, \ldots, y_{It})^T$ denote the crop yield at $ith$ location in year $t$, $i = 1, \ldots, I$, $N_i$ is the set of neighbors of location i. Locations $i = 1, \ldots, n$ forms a lattice. $n_i$ is the number of those neighbors of location i. The spatial dependence or spatial correlation is measured through the parameter $\rho$.[1] The mean of $Y_{it}$ can be modeled as functions of the fitted residuals from neighboring plots. The rationale is that if the plots that are neighbors of plot i have negative residuals indicating that they are of below–average yields for this county then county $i$ is likely to be below average in average yield as well. This suggests using the average of residuals from a county's neighbors as a factor to explain the mean of $Y_{it}$.

The AR model requires structure on the spatial smoothing. One way to specify the shape of the regions is to use a neighborhood matrix **W** which indicates whether the regions touch or not. The rows and columns of this matrix correspond to the observations. We assume the element in the matrix **W** to be equal to 1 if two regions are neighbors and 0 if the two regions are not neighbors. Define $W = (w_{ij})$, where $W$ is an $n \times n$ matrix whose nonzero elements specify the neighboring times of each yield data point:

$$w_{ij} = \begin{cases} 1 & \text{if region } i \text{ shares a common edge or border with region } j, \\ 0 & \text{if i=j}, \\ 0 & \text{otherwise} \end{cases}$$

By using the notation of the neighborhood matrix, the conditional distribution of $y_{it}$ is as follows:

$$y_{it}|y_{j\neq i,t} \sim N(\mu_{it} + \rho_t \sum_{j\in N_i} w_{ij}(y_{jt} - \mu_{jt}), \frac{\sigma_t^2}{n_i}) \tag{2}$$

where $\mu_{it} = \beta_{0i} + \beta_{1i}t$. $\mu_{it}$ can also be modeled more generally by using non-parametric regression method. The conditional mean of $E(y_{it}|y_{j,t}) = \mu_{i,t} + \rho_t \sum_{j\in N_i}(y_{j,t} - \mu_{j,t})$, and the conditional variance is $\frac{\sigma_t^2}{n_i}$. $\rho_t$ is referred to as 'spatial correlation or spatial dependence' parameter and controls the strength of spatial association. $\sum_{j\in N_i} w_{ij}(y_{jt} - \mu_{jt})$ is a weighted

[1] $\rho$ can be explained as the spatial correlation between neighbor counties.

average of the spatial effects for all neighbors of location i other than itself.

By Brook's lemma, the joint distribution of $Y_t$ in matrix notation is as follows:

$$Y_t = (y_{1t}, \ldots, y_{It})' \sim (\mu_t, \sigma_t^2 (I - \rho_t * w)^{-1} D) \tag{3}$$

where $D = diag(\frac{1}{n_1}, \ldots, \frac{1}{n_n})$ and $n_i = \sum_{j \neq i} w_{ij}$ is the number of neighbors at each location $S_i$.

The AR model specified in this way with a single parameter times some weighting matrix has been used extensively for modeling irregular lattices in applied econometrics. (e.g. Anselin and Florax (1995), Kelejian and Prucha (1999), and Anselin(2002)). The maximum likelihood estimation method can be carried out to estimate the model.

The likelihood function $L(\theta)$ is defined as follows. The parameters in this model are $\theta = (\beta_{01}, \ldots, \beta_{0I}, \beta_{11}, \ldots, \beta_{1I}, \beta_{21}, \ldots, \beta_{2I}, \sigma_1^2, \ldots, \sigma_T^2, \rho_1, \ldots, \rho_T)$.

$$L = |\sigma_t^2 (I - \rho_t w)^{-1} D|^{-\frac{1}{2}} \exp(-\frac{1}{2}(Y_t - \mu_t)'(\sigma_t^2 (I - \rho_t w)^{-1} D)^{-1}(Y_t - \mu_t))$$

where $\mu_t = (\mu_{1t}, \ldots, \mu_{It})'$, $D = diag(\frac{1}{n_1}, \ldots, \frac{1}{n_I})$ and $n_i = \sum_{j \neq i} w_{ij}$. Then the log-likelihood function $LLF(\theta)$ is:

$$LLF = \sum_t (\log \sigma_t^2 - \log |I - \rho_t w| + \frac{1}{\sigma_t^2}(Y_t - \mu_t)' D^{\frac{1}{2}} (I - \rho_t w) D^{\frac{1}{2}} (Y_t - \mu_t))$$

## 2.3 Directional Spatial Autoregressive Model

One of the limitations of the AR model is that the neighbors are formed using a neighborhood matrix and it assume the same spatial correlation in all directions by assigning equal weight to all directions. This is counter-intuitive for the spatial correlation of crop yields. Usually, due to the bigger weather similarity of West-East (W-E) direction than South-North (S-N) direction, the yields distribution between the west and east should be more similar than that of the north-south. Thus, we can expect that the spatial correlation between W-E should be bigger than that of S-N. In order to incorporate the idea of directional effects in the spatial correlation, a directional AR model (DAR) are proposed in this study. In some recent literature, there have been some attempts to use different CAR models for different parts of the region. White and Ghosh (2008) develop a CAR model with stochastic parameters to determine effects of the neighborhood. Kyung and Ghosh (2008) present a directional CAR (DCAR) model to accommodate spatial variations by using different weights to neighbors in different directions in a Bayesian framework.

Here we adopt the Kyung and Ghosh (2008)'s DCAR approach to estimate the spatial correlation of crop yields in different regions in the U.S. and the estimated spatial correlation will make implications to price group insurance plan such as Group Risk Plan (GRP).

The subset based on different directions are denoted as $N_{ik}$, where $k$ stands for different direction. For example, if there are two directions (S-N and W-E), $N_{i1}$ and $N_{i2}$ can be constructed based on the associated S-N and W-E neighborhoods. The directional neighborhood matrices are $W^{(1)} = ((\omega_{ij}^{1}))$ and $W^{(2)} = ((\omega_{ij}^{2}))$, respectively. That is,

$$w_{ij}^{k} = \begin{cases} 1 & \text{if region } i \text{ shares a common edge or border with region } N_{ik}, \\ 0 & \text{if i=j}, \\ 0 & \text{otherwise} \end{cases}$$

Notice that $W = W^{(1)} + W^{(2)}$ are the same neighborhood matrix as in the regular model.

Based on the directional neighborhood matrix $W$, a DAR model can be developed to account for anisotropy.

Let $\rho_t^{(1)}$ and $\rho_t^{(2)}$ denote the directional spatial effects corresponding to $N_{i1}$ and $N_{i2}$. $\rho_t^{(1)}$ and $\rho_t^{(2)}$ are the S-N and W-E spatial correlation respectively. The distribution of $Y(S_i)$ conditional on the rest of $Y_t$ can be expressed based on the first two moments:

where $\omega_{ij}^{(k)} \geq 0$ and $\omega_{ii}^{k} = 0$ for $k = 1, 2$ and $m_i = \sum_{k=1}^{2} \sum_{j=1}^{n} \omega_{ij}$

The directional spatial correlations are defined as $\rho_t^{(1)}$ and $\rho_t^{(2)}$. $\rho_t^{(1)}$ and $\rho_t^{(2)}$ are the S-N and W-E spatial correlation respectively.

The joint distribution of $Y$ is:

$$Y_t = (y_{1t}, \ldots, y_{It})' \sim N(\mu_t, \sigma_t^2 (I - \rho_t^{(1)} w^{(1)} - \rho_t^{(2)} w^{(2)})^{-1} D), t = 1, \ldots, T \quad (4)$$

The likelihood function $L(\theta)$is:

$$L = |\sigma_t^2 (I - \rho_t^{(1)} w^{(1)} - \rho_t^{(2)} w^{(2)})^{-1} D|^{-\frac{1}{2}} \exp(-\frac{1}{2}(Y_t - \mu_t)' (\sigma_t^2 (I - \rho_t^{(1)} w^{(1)} - \rho_t^{(2)} w^{(2)})^{-1} D)^{-1} (Y_t - \mu_t))$$

where $\theta = (\beta_{01}, \ldots, \beta_{0I}, \beta_{11}, \ldots, \beta_{1I}, \beta_{21}, \ldots, \beta_{2I}, \sigma_1^2, \ldots, \sigma_T^2, \rho_1^{(1)}, \ldots, \rho_T^{(1)}, \rho_1^{(2)}, \ldots, \rho_T^{(2)})$. The number of parameters to be estimated is $3 * I + 3 * T$ with sample size equal to $I * T$. We need $I * T > 3 * I + 3 * T$ which is satisfied if $I > 6$ and $T > 6$. To avoid large dimensional optimization for $\theta$, we estimate the MLE for the spatial parameters at each time points $t$. The optimization is done by using the optim package in R.

The log-likelihood function $LLF(\theta)$ is:

$$LLF = \sum_t (\log \sigma_t^2 - \log |I - \rho_t^{(1)} w^{(1)} - \rho_t^{(2)} w^{(2)}| + \frac{1}{\sigma_t^2}(Y_t - \mu_t)' D^{\frac{1}{2}} (I - \rho_t^{(1)} w^{(1)} - \rho_t^{(2)} w^{(2)}) D^{\frac{1}{2}} (Y_t - \mu_t))$$

# 3 Data

The county-level corn yields data from Iowa are used in this analysis. The lattice is formed by the counties of a Iowa. The centroid data of each county is used to calculate the direction of the spatial correlation. Iowa is in the central Corn belt where climate and soil are ideal for corn production, so the corn yield volatility will tend to be low in this state. 99 counties Iowa all-practice corn yield data are selected from the time period of 1926-2007. By using the CAR model, the state level average yields is obtained which will take the spatial correlation into consideration. The same methodology can be used to obtain the county-level yield estimate if the lattice of farm-level data is available.

# 4 Estimation Results and Economic Implications

The trends of yields are found to be different across space (figure 1 and 2). The estimates of the spatial correlation are found to be varying in time with an average spatial correlation equal to 0.17 by using autoregression model. If the directional effect is taken into account, the spatial correlation of west-east (W-E) is greater than the spatial correlation of north-south (N-S) (0.19 vs. 0.15 in average) as shown in figure 3, which is consistent with the agricultural production practice. The directional effect is expected to be more significant for larger geographic units. The different spatial correlations in directions justify the need of using directional effects in modeling the spatial correlation of yields.

Moreover, the pairwise spatial correlation between any two different location $S_i$ and $S_j$ can be obtained with the information of neighborhood weighting matrix $w_{ij}$ and the estimated $\rho$, although the estimated spatial correlation $\rho$ in the AR model does not have a linear relationship with the implied neighbor correlation. Implications for optimal area-yield insurance program can be drawn from the estimates of spatial correlation and any pairwise spatial correlation analysis. If we want to design crop insurance contract with large spatial regions, this DAR model is a powerful tool to capture spatial trends that may be presented in the data. The estimated distance and directions in which the area crop yields from the state are significantly spatially correlated will give implications to the range and the shape of the geographical area to be considered in measuring yield risk and design of reinsurance and the area-yield index for the GRP. The implied spatial correlation between different counties should help accurately measure the systemic yield risk and thus improve the efficiency of risk management.

An extension of the spatial analysis will go to the spatial correlation for different crops, e.g.,corn and soybeans, by using a copula approach in the DAR model. The county-level corn and soybean yield data from Iowa state can be used in this analysis. The bivariate distribution obtained in the copula analysis allows for an analysis of the higher-order spatial relationships and higher-order cross-moments of crop yields. These can be used in the design of area-yield insurance contracts that account for the spatial correlation of different crop yields. The questions addressed in the spatial analysis for bivariate crop yields include whether it is appropriate to use one bivariate distribution, a Clayton copula as an example, over a certain geographical area for corn and soybean yields; and whether the decay rate of the spatial correlation in bivariate case is similar to that in univariate case. If not, the pattern of the change in the parameters of bivariate distribution needs to be captured by the spatial analysis to make policy implication to the optimal shape and size of geographical area in the combination area-yield insurance design.

# References

[Anselin(2002)] Anselin, L. 2002. "Under the hood Issues in the Specification and Interpretation of Spatial Regression Models." *Agricultural Economics* 27:247–267.

[Anselin and Florax(1995)] Anselin, L., and R. Florax. 1995. *New Directions in Spatial Econometrics*. Springer, Berlin.

[Cressie(1993)] Cressie, N. 1993. *Statistics for Spatial Data*. John Wiley & Sons, New York.

[Goodwin and Ker(1998)] Goodwin, B.K., and A.P. Ker. 1998. "Nonparametric Estimation of Crop Yield Distributions: Implications for Rating Group-Risk(GRP) Crop Insurance Contracts." *American Journal of Agricultural Economics* 80:139–53.

[Kelejian and Prucha(1999)] Kelejian, H., and I. Prucha. 1999. "A Generalized moments Estimator for the Autoregressive Parameter in a Spatial Model." *International Economic Review* 40(2):509–533.

[Ozaki et al.(2008)Ozaki, Ghosh, Goodwin, and Shirota] Ozaki, V.A., S.K. Ghosh, B.K. Goodwin, and R. Shirota. 2008. "Spatio-Temporal Modeling of Agricultural Yield Data with an Application to Pricing Crop Insurance Contracts." *American Journal of Agricultural Economics* 90(4):951–961.

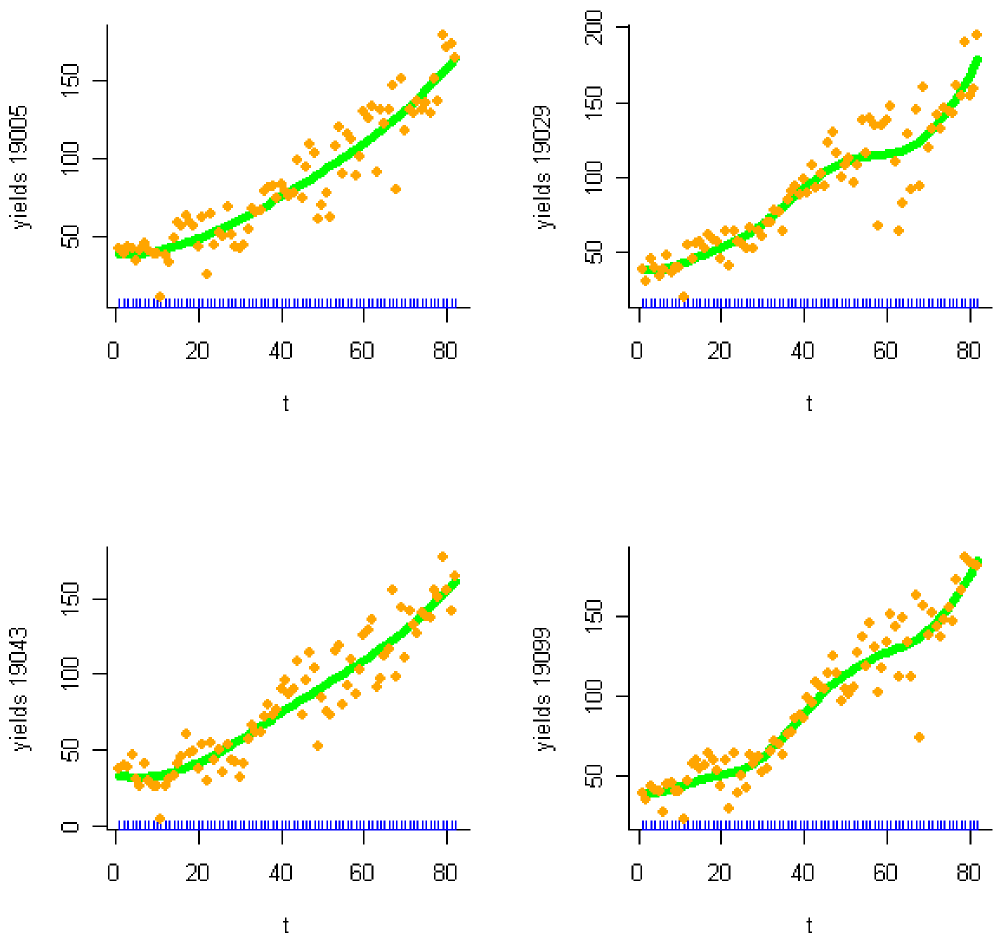


Figure 1: Scatter Plot of Corn Yields v.s. Time—Selected Counties in Iowa

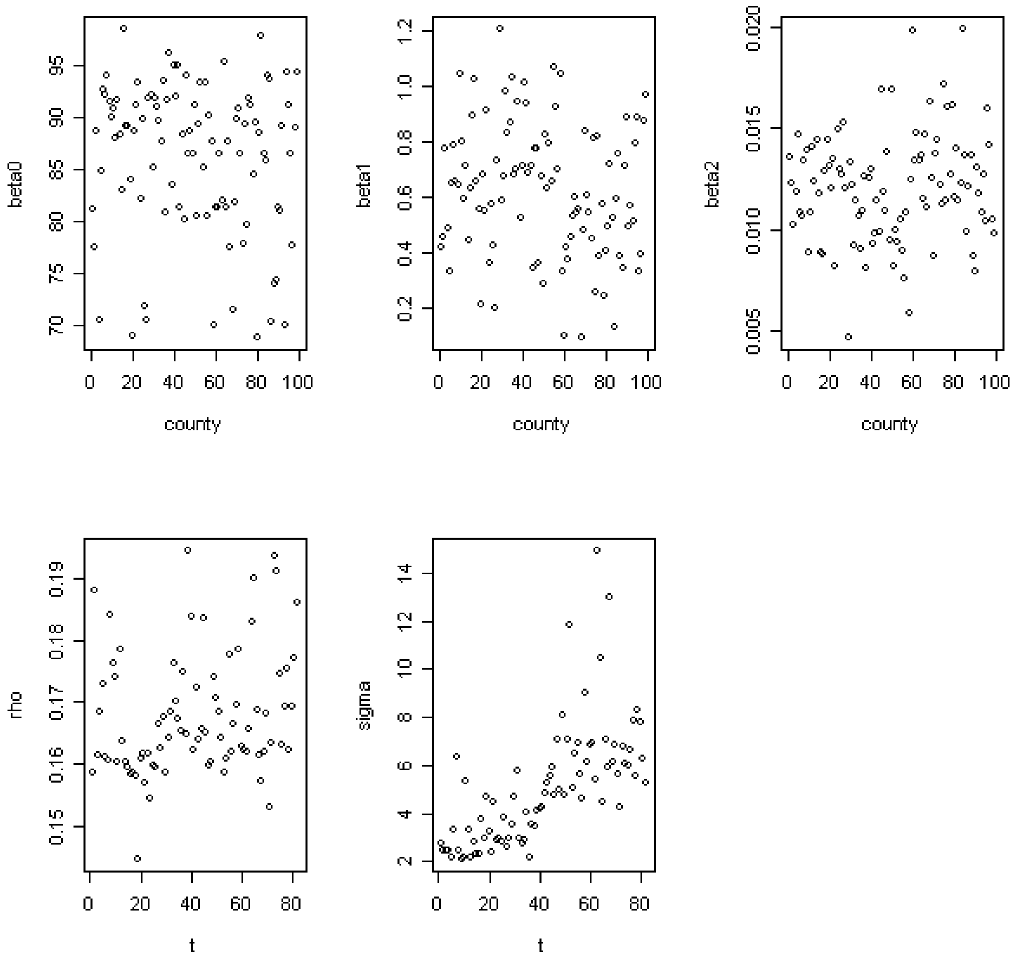


Figure 2: Space Time Plots of Regression Coefficients

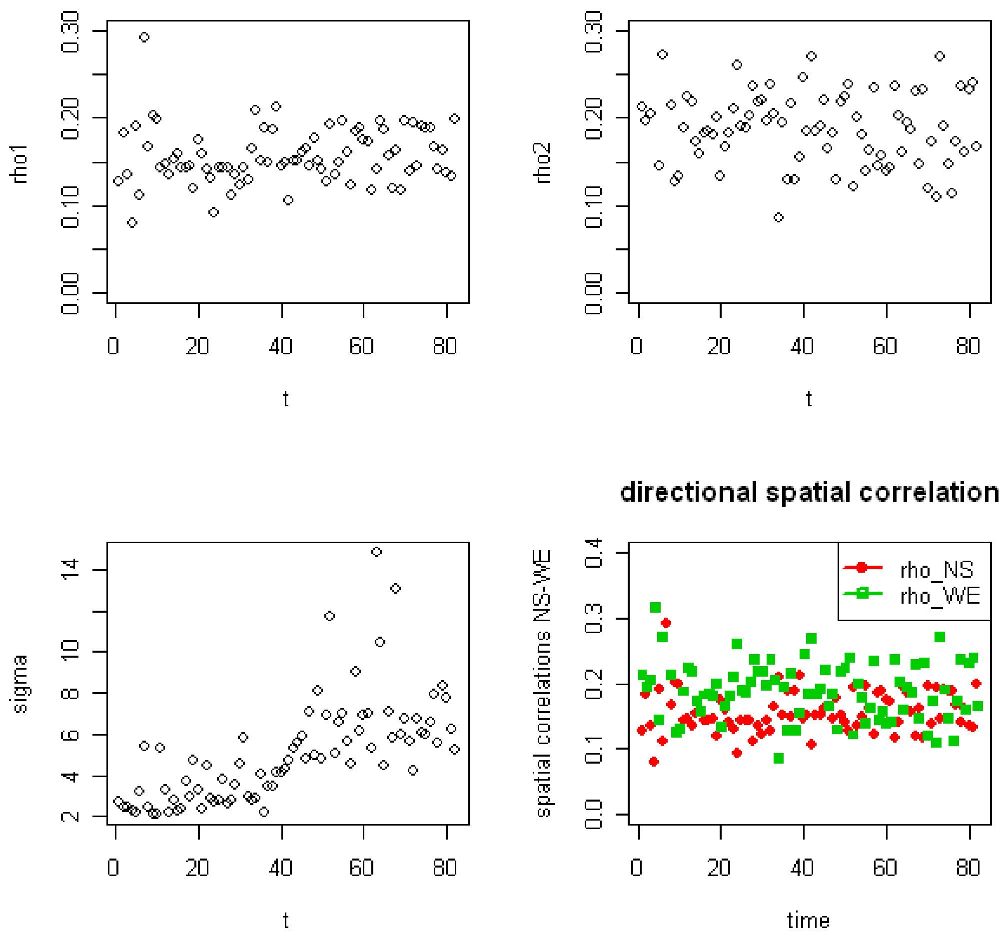


Figure 3: Temporal Plot of Directional Autoregression Model