COMPARISON OF TWO REGRESSION MODELS FOR PREDICTING CROP YIELD

Li Zhang[1], Lei Ji[2], Liping Lei[1,] and Dongmei Yan[1]

[1]Center for Earth Observation and Digital Earth, Chinese Academy of Sciences
[2] ASRC Research and Technology Solutions, contractor to the U.S. Geological Survey (USGS) Earth Resources Observation and Science (EROS) Center

1. INTRODUCTION

Empirical regression models have been developed for crop yield predicting. NDVI alone or coupling with other environmental and climatic data sets has been successfully used to predict crop yield. Linear regression models are the most common methods in these studies. But the conventional regression technique based on the ordinary least square (OLS) estimation are not adequate, because possible spatial autocorrelation among variables violates the underlying assumption that observations are independent, which may lead to a biased estimation of the standard errors of model parameters and mislead significance test [1-2]. Such regression models may mistakenly emphasize on some independent variables that in fact have little or no influence on the dependent variable. To properly identify the relationship between dependent and independent variables, it is important to adjust for autocorrelation in regression analyses that are associated with geographic data [2]. Thus, a spatial regression technique that attempts to correct for spatial autocorrelation in the regression model is required.

An appropriate quantification of the relationship of crop yield to NDVI and climatic data is critical for predicting crop yield. This study tried to answer two questions: 1) which environmental variables have the significant affect on crop yield using the OLS regression model and spatial autoregressive model and 2) how those environmental variables behavior differently in predicting crop yield in dry and normal years.

2. STUDY AREA AND DATA

The study covered 99 counties of Iowa State located in the west Corn Belts of U.S. The county-level annual corn yield data (corn for grain) for 2003 and 2004 were obtained from the USDA National Agricultural Statistics Service database. The NDVI data were derived from MODIS at 1-km resolution and 16-day intervals. The NDVI analysis was spatially restricted to Row Crops class indicated by the 30-m National Land Cover Dataset (NLCD 1992). The NDVI was averaged for the growing season (June through August) for years 2003 and 2004 and for each county. Precipitation and temperature were acquired from the NOAA Climate Prediction Center (CMAP). Mean temperature and total precipitation were calculated for the period from June to August in 2003 and 2004. Water holding capacity (WHC) data were derived from the State Soil Geographic (STATSGO) database.

3. METHODS

3.1 Ordinary least squares (OLS) regression analysis

The relationship between corn yield and environmental variables was quantified using the OLS regression model, which does not account for spatial autocorrelation.

3.2 Assessing spatial autocorrelation of variables

Environmental data generally show spatial dependence, thus they violate the assumption of independent observations. The global Moran's I using Monte-Carlo simulation and Moran local indicators of spatial association (LISA) were used to test for global and local spatial clustering, respectively. The LISA map provides an indication of the extent of significant spatial autocorrelation and the intensity of autocorrelation [3].

3.3 Spatial autoregressive analysis

When spatial autocorrelation exists, an alternative is to use spatial autoregressive model to adjust for spatial autocorrelation inherent in these data. The spatial autoregressive models use the geographic weights matrix coupled with a spatial autocorrelation parameter, ρ, to account for spatial autocorrelation by filtering it from the georeferenced data during model parameter estimation, and by which the accompanying residuals are spatially independent [4]. In this study, the spatial lag model was adopted based on the Lagrange Multiplier (LM) test.

3.4 Evaluating of model performance

The performance of the OLS and spatial autoregressive models were spatially assessed in terms of spatial distribution and Moran coefficient of model residuals. Spatial correlograms of model residuals were also generated using Moran's I coefficients.

4. RESULTS

4.1 Dry year of 2003

4.1.1 OLS model

The OLS model fitted the data well with $R^2 = 0.79$ and standard error (SE) = 8.138. All the four variables were statistically significant ($p < 0.05$). A significant positive relationship was observed between corn yield and independent variables (NDVI, precipitation, WHC), and a negative relationship was observed between corn yield and temperature in the dry year of 2003.

4.1.2 Assessing spatial autocorrelation of variables

The LISA map indicated that the distribution of corn yield, NDVI, and other environmental variables are spatially autocorrelated, which violates the underlying assumption of regression analysis. The corn yield and NDVI LISA map showed a higher cluster in the south. Consequently, OLS model tended to underestimate SE of model parameters, and inferences about corn yield on NDVI and other environmental variables may be misleading, and thus overstated the role of some environmental predictors.

4.1.3 Spatial autocorrelation Analysis

The comparison between spatial lag model and OLS mode showed that temperature and WHC had become insignificant variables in the spatial lag model. The R^2 of lag model increased 11% over the OLS model. Temperature and WHC were removed from the spatial lag model, one variable at a time. The reduced spatial lag model were significant with $R^2 = 0.88$.

4.2 Normal year of 2004

4.2.1 OLS model

It showed that corn yield was positively correlated to NDVI and WHC, and negatively correlated to precipitation, but not significantly correlated to temperature. It is noticed that total precipitation is positively correlated with corn yield for dry year of 2003, but negatively correlated with corn yield for normal year of 2004.

4.2.2 Spatial autocorrelation analysis

Precipitation was not significant with $p = 0.09$. As a result, precipitation was dropped from the spatial lag model and only NDVI was remained. The R^2 had increased from 0.50 in OLS model to 0.58 in spatial lag model. As seen from above analysis for 2003 and 2004, autoregressive model fit the data better than the OLS model. The spatial lag model improved the model performance by increasing R^2 and reducing SE and AIC.

4.3 Spatial assessment of model residuals

For both 2003 and 2004, the OLS model produced larger residuals than spatial lag model. The residuals from spatial lag model showed less spatial autocorrelation than that from OLS model. The spatial correlograms (Figure 1) of residuals from OLS model showed significant autocorrelation up to lags of 2, and Moran's I ranges from -0.2 to 0.5. However, there was no significant spatial autocorrelation for residuals from the spatial lag model, and the residuals ranges from -0.15 to 0.15. Moran's I from the OLS model residuals confirmed that the assumption of independently distributed errors was violated. Lower spatial autocorrelation was found in the lag model residuals, suggesting the capability of lag model to deal with spatial non-stationary problems.

5. DISCUSSIONS AND CONCLUSIONS

Spatial lag model indicated a significant improvement in model performance over OLS model as indicated by higher R^2, lower AIC, and lower SE. Autoregressive model is capable of adjusting the spatial autocorrelation in model residuals, which is ignored by the OLS model.

The study demonstrated that NDVI, precipitation and WHC are significant predictors on corn yield in dry year, but NDVI is the only significant predictor in normal year. High precipitation in the normal year is not a major determinant for corn yield. Models performed better for forecasting corn yield during dry year than during normal

years. It is consistent with the study by Mkhabela et al. [5] that there is relatively poor correlation between maize yield and NDVI in the high rainfall year.

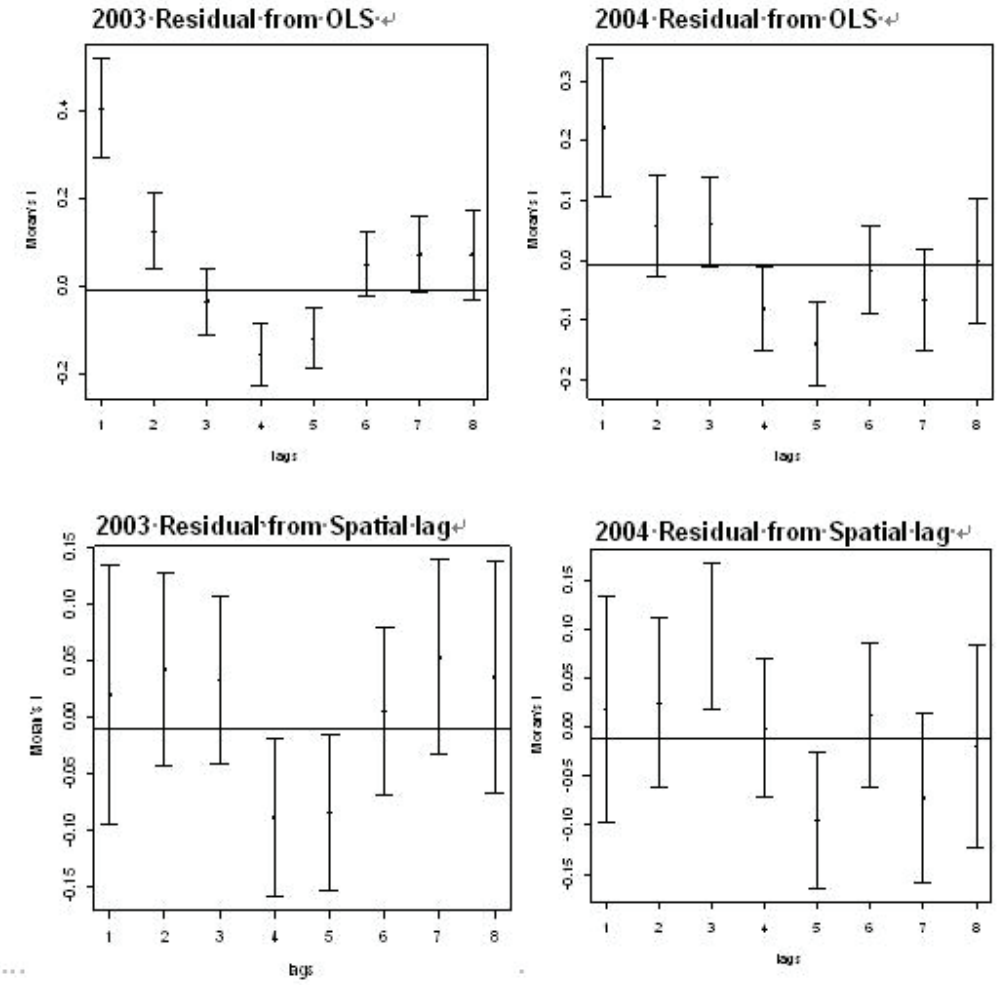

Figure 1 Spatial correlograms for model residuals

6. REFERENCES

[1] L. Anselin and D.A. Griffith, "Do spatial effects really matter in regression analysis?," *Papers in Regional Science*, 1988, 65, 11-34.

[2] D.A.Griffith, "Introduction: the need for spatial statistics," In S.L. Arlinghaus and D.A. Griffith (Ed.), *Practical Handbook of Spatial Statistics,* CRC Press, Boca Raton, 1996, pp. 1–15.

[3] L. Anselin, "Local indicators of spatial association – LISA," *Geographical Analysis,* 1995, 27(2), 93–115.

[4] L. Anselin,. *Spatial Econometrics: Methods and Models,* Kluwer Academic Publishers, Dordrecht, 1988.

[5] M.S.Mkhabela, M.S. Mkhabela,, and N.N.Mashinini, "Early maize yield forecasting in the four agro-ecological regions of Swaziland using NDVI data derived from NOAA's-AVHRR," *Agricultural and Forest Meteorology,* 2005, 129(1-2), 1-9.