Improving Crop Yields Forecasting Using Weather Data: A Comprehensive Approach Combining Principal Component Analysis and Credibility Model

Extended Abstract

Wenjun Zhu[†]

Lysa Porth[‡]

Ken Seng Tan[†]

[†] Department of Statistics and Actuarial Science, University of Waterloo

‡ Department of Agribusiness and Agricultural Economics, University of Manitoba

Background

Crop insurance premium rating can be quite challenging, largely due to extreme weather events that can often cause widespread losses across many geographic regions. Further, evidence suggests that adverse weather, such as droughts and floods, may be becoming more prevalent, and these changes together with diminishing returns from technological advances may limit the ability of many regions to achieve the necessary gains (Cassman, 1999; Lobell and Asner, 2003; Dai et al., 2004), which augments ratemaking difficulties for government crop insurance agencies, as well as private sector insurers and reinsurers.

Yield models are an integral component to ratemaking algorithms, the prediction of which is essential crop insurers and reinsurers. However, the natural systems are complex, and the crop yield is an unified bio-socio-system comprised of complex interaction among the soil, the air, the water, and the crops grown in it, where a comprehensive model is required which are possible only through classical engineering expertise (Murthy, 2004). In the statistical modeling framework, a lot of simplifications will be embodied in sacrifice for the feasibility and efficiency. The most investigated statistical crop-yield-weather models are multivariate regression models (Alexandrova and Hoogenboom, 2000; Prasada et al., 2006; Yu, 2011). However, considering the inherent and irreparable disadvantages of the multiple regression model, a more scientific methodology to incorporate weather data into crop yield models, is still under exploration, and remains of great importance to government, and private sector insurers, and reinsurers.

The objective of this paper is to develop an improved crop forecasting model, through the incorporation of weather data. A new credibility model is developed by combining Principal Component Analysis (PCA), semi-linear credibility, and regression credibility models. This

novel approach is capable of aggregating vast quantities of information from individual local weather stations data, and constructs various possible weather indexes representing weather conditions across the whole geographical region. Empirical results show that this new model is able to provide better in-sample and out-of-sample yield forecasting results.

Data summary

The yield data set used in this paper is a farm level yield data of Manitoba for 216 types of crops from 1996 to 2011. The data panel covers totally 19238 farms from 125 municipalities (or equal regions). We aggregate the yields according to their acres for farms in the same municipality in order to eliminate the noise. Given the fact that farmers usually alternate their crops, which helps to keep the soil healthy and fertile, yields of most individual crop cannot provide satisfying complete time series for analysis and modeling. Therefore, in our analysis, we weight each crop by the acre and create the new "Weighted Yield" as the basic random variable in the following analysis. After deleting the 6 municipalities whose "Weighted Yield" series contains zeros, there are finally 119 municipalities entering the empirical analysis.

The original weather data set analyzed includes daily temperature (maximum, minimum, and mean) from 24 weather sations in Manitoba, daily precipitation from 30 weather stations in Manitoba, and Palmer Drought Severity Index (PDSI) for Manitoba. The locations of the weather stations are available, The temperature and precipitation data are publicized Adjusted and Homogenized Canadian Climate Data (AHCCD) from the website of Environment Canada, and the earliest records go back to 1875. The Palmer Drought Severity Index (PDSI) data analyzed in this article are downloaded from website of National Center for Atmospheric Research (NCAR) Earth System Laboratory Climate & Global Dynamics

(CGD) Climate Analysis Section, where a global PDSI from 1850 to 2010 is available.

In the temperature and precipitation data, at some timepoints, there are some weather stations that have missing data, which obstructs the further analysis. There are a number of alternative ways to deal with missing data, such as permutation and interpolation. In this article, we appeal to a method called Kriging to predict the missing data.

In order to extract as much information from the data we have as possible, we construct a weather index system containing 28 weather variables from the daily temperature and precipitation data together with the monthly PDSI data. In addition to considering the annual and growing season weather data, we also referred to the heat unit system widely used in agrology (Suomi and Newman, 1960;Brown, 1969). We developed the three kinds of heat indexes, the definition of which is according to the Corn Heat Units (CHU). Developed by Brown in 1969 to apply in Ontario, CHU concept is now widely used across Canada. The parameters in them are chosen according to paper by Brown and BootsmaBrown and Bootsma (1993).

Model Construction

The model proposed in this article is **Principal Component Analysis-Credibility Model (PCACM)**, which compresses Principal Component Analysis (PCA) and credibility model into the original regression model. The starting point of the attempt to apply credibility approach is the fact that the crop production risk is usually geographically correlated and highly variable, and ignoring the spatial correlation will lead to either overestimation or underestimation of the crop yields, which will finally cause crop insurance market failures

(Glauber and Collins, 2004; Woodard et al., 2012). Credibility estimator of the crop yield in PCACM (expressed as in Equation 0.1) is able to minimized the Quadratic Loss of the credibility estimator and give improved in-sample and out-of-sample forecasting results.

$$\widehat{\mu_{\boldsymbol{X}}(\Theta)} = \boldsymbol{\mu_X} + \frac{\tau_{\boldsymbol{XY}}}{\tau^2_{\boldsymbol{Y}}}[(W - \boldsymbol{\mu})'\Gamma]_{n\times p} A(\boldsymbol{B} - \boldsymbol{\beta}), \tag{0.1}$$

Forecasting Results

Table 1 summarizes the statistic characteristics of the distribution R^2, AIC, and standard errors (SE) of regression coefficients for 119 municipalities under both multiple regression model and PCA regression model. More directly, Figure 1 and Figure 2 are the histograms of the R^2 values of the regression model and PCA regression model respectively. Obviously, for most municipalities, the PCARM R^2 are larger than 0.9, while the R^2 of more than half of the multiple regression models are below 0.5, which shows a higher explanatory capacity of PCARM compared to the original regression model. Although containing more parameters to estimate, the PCARM still has generally lower AIC indicating a better goodness of fit. Moreover, we can see that the average standard errors of the coefficients of PCARM are super lower than the multiple regression model, meaning the PCARM is more robust than the multiple regression model.

Provided that the better fitting results of PCARM than the original regression model, we are expecting better forecasting abilities of the PCARM as well as the PCACM, which combines credibility idea in the model. Table 2 listed the statistical summary of both in-sample and out-of-sample forecasting errors of the three models. We can see that the PCARM and PCACM have competitive performances in in-sample-forecasting, both of which are vastly

	Original Regression Model			PCA Regression Model (PCARM)		
	R^2	AIC	SE($\times 10^5$)	R^2	AIC	SE($\times 10^5$)
Mean	0.6553	369.79	0.9149	0.9373	349.76	0.00018
Variance	0.0617	1280.57	1093000	0.0019	925.53	0.00215
25%	0.4344	346.12	0.00256	0.9208	330.67	0.00007
50%	0.5802	374.63	0.01174	0.9476	354.87	0.00015
75%	0.9027	398.84	0.07245	0.9688	372.62	0.00025

Table 1: Statistical Summary of Results from Two Models

	In-Sample Errors ($\times 10^6$)			Out-of-Sample Errors ($\times 10^6$)		
	Regression	PCARM	PCACM	Regression	PCARM	PCACM
Mean	2608.46	435.13	471.82	74635.57	9112.83	552.22
SE	5088.05	781.38	761.63	159681.00	15750.29	919.15
0%	2.21	0.45	47.15	39.93	11.77	46.10
25%	172.91	37.50	86.34	3134.02	927.64	88.74
50%	828.88	172.85	222.23	15736.79	3815.40	230.47
75%	3050.89	515.94	557.89	61280.14	11198.33	656.02
100%	35982.25	4978.88	4894.11	994106.00	101695.00	5928.30

Table 2: Statistical Summary of Results from Two Models

superior than the in-sample performance of original regression model. For the out-of-sample forecasting, we can see that the PCARM is better than regression, while the PCACM, by combining PCA regression with credibility approach, is able to provide much better out-of-sample forecasting results.

Figure 1: Pie plot indicating the weather index left in the final regression model for all municipalities.

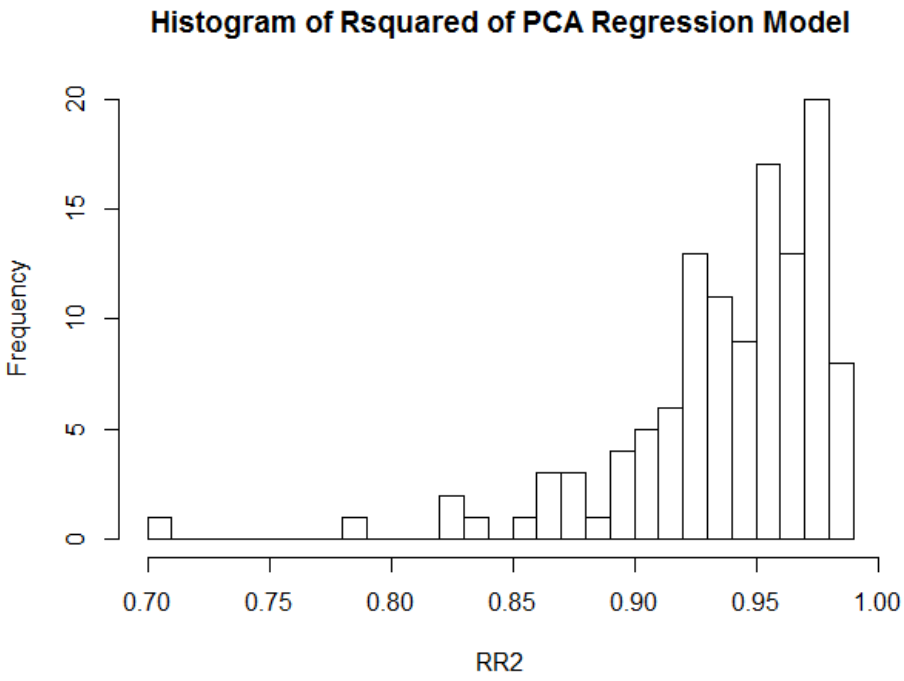

Figure 2: Pie plot indicating the weather index left in the final regression model for all municipalities.

References

V. A. Alexandrova and G. Hoogenboom. The Impact of Climate Variability and Change on Crop Yield in Bulgaria. *Agricultural and Forest Meteorology*, 104:315–327, 2000.

D.M. Brown. Heat units for corn insouthern Ontario. *Ontario Department of Agriculture and Food, Toronto, Ontario Information Leaflet*, pages 111–131, 1969.

D.M. Brown and A. Bootsma. Crop heat units for corn and other warm season crops in Ontario. *Earsel Advances in Remote Sensing, Ontario Ministry of Agriculture and Food Factsheet*, pages 93–119, 1993.

Envrionment Canada. Adjusted and Homogenized Canadian Climate Data (AHCCD). `http://ec.gc.ca/dccha-ahccd/default.asp?lang=En&n=B1F8423A-1`.

Kenneth G. Cassman. Ecological intensification of cereal production systems: Yield potential, soil quality, and precision agriculture. *Proceedngs of the National Academy of Science of the United States of America*, 96(11):5952–5959, 1999.

A. Dai, K. E. Trenberth, and T. Qian. A global data set of Palmer Drought Severity Index for 1870-2002: Relationship with soil moisture and effects of surface warming. *J. Hydrometeorology*, 5:1117–1130, 2004.

National Center for Atmospheric Research (NCAR) Earth System Laboratory Climate & Global Dynamics (CGD) Climate Analysis Section. Palmer Drought Severity Index (PDSI). `http://www.cgd.ucar.edu/cas/catalog/climind/pdsi.html`.

J. W. Glauber and K. J. Collins. Crop Insurance, Disaster Assistance, and the Role of the Federal Government in Providing Catastrophic Risk Protection. *Agriculture Finance Review*, 62(2):80–101, 2004.

David B. Lobell and Gregory P. Asner. Climate and Management Contributions to Recent Trends in U.S. Agricultural Yields. *Science*, 299:1032, 2003.

V. Radha Krishna Murthy. Ecological intensification of cereal production systems: Yield potential, soil quality, and precision agriculture. In *Satellite Remote Sensing and GIS Applications in Agricultural Meteorology*, volume AGM-8, 2004.

Anup K. Prasada, Lim Chai, Ramesh P. Singha, and Menas Kafatos. Crop Yield Estimation Model for Iowa Using Remote Sensing and Surface Parameters. *International Journal of Applied Earth Observation and Geoinformation*, 8:26–33, 2006.

V.E. Suomi and J.E. Newman. A critique of the heat unit approach to plant response studies. *Ecology*, 41:785–790, 1960.

Joshua D. Woodard, Gary D. Schnitkey, Bruce J. Sherrick, Nancy Lozanno-Gracia, and Luc Anselin. A Spatial Econometric Analysis of Loss Experience in the U.S. Crop Insurance Program. *Journal of Risk and Insurance*, 79(1):261–285, 2012.

Tian Yu. *Three Essays on Weather and Crop Yield*. PhD thesis, Iowa State Univeristy, 2011.