Use of dynamic linear model for predicting crop yield trends in foresight studies on food Security

David Makowski

Lucie Michel

UMR 211 INRA AgroParisTech
BP01
Thiverval-Grignon, France
makowski@grignon.inra.fr

ABSTRACT

The world's population is projected to pass 9 billion in 2050 and there is increasing concern about the capability of agriculture to feed the world. Foresight studies on food security require predictions of future yield trends (crop production per unit of soil area). These trends are usually estimated from yield time series provided by statistical agencies at national and regional scales. Different types of statistical methods have been proposed to analyze yield time series (linear regression, non linear regression, moving average etc.) but, so far, the predictive performances of these methods have not been evaluated. This paper will describe a simple space-state model to analyze yield time series for several major crops in the world (e.g., wheat, maize, rice). The proposed model is a dynamic linear regression model predicting future yield trends and their associated credibility intervals using the Kalman filter algorithm. The accuracy of yield predictions obtained with our method is evaluated using wheat yield data provided by the Food and Agriculture Organization of the United Nations. We show with these data that the dynamic linear regression model is more flexible and performs better than most of the statistical methods currently used to analyze yield time series

Keywords: Dynamic linear model; Kalman filter; Time series; Yield trend

1. Introduction

Several studies have recently shown that, after a period of strong yield increase, yield levels are currently stagnating in several countries. The yield of cereal crops increased by more than 98% worldwide, and by more than 187% in France from 1960 to 1990 (FAOSTAT). However, since the 1990s, yield increases for several major cereal crops (wheat, maize, rice, barley or oat) have slowed down. In some countries, yield levels have remained constant or have even declined for some crops (Brisson et al., 2010; Finger, 2010). These results have raised significant concern in the scientific community about the ability of agriculture to feed the world in the future.

Statistical analyses play a key role in current research studies on food security where yield time series analysis is used to estimate past yield trends and to predict future yield trends. Various types of statistical models have been used for the analysis of yield time series. Linear regression has been used in many studies. Other regression models, such as quadratic regression, bi-linear, tri-linear, and linear-plus-plateau models, have been used in a smaller number of papers (Brisson et al., 2010; Finger, 2010). Several authors have shown that quadratic and linear-plus-plateau models tend to perform better in cases of yield stagnation. Statistical methods other than regression models have been used to predict future yield trends. Kumar (2000) compared the performances of linear and quadratic regression models with those of exponential smoothing (also known as the Holt-Winters method) and moving averages. However, this result was obtained with a small dataset: yield predictions were assessed for three years at a specific location in Canada. It is, thus, difficult to draw general conclusions from this study.

The dynamic linear model (DLM) is a recently developed statistical method that can be used to estimate past trends and to predict future trends in time series (Petris et al., 2009). This method is based on the Kalman filter and the Kalman smoother. It is very flexible, because the coefficients of the underlying linear model are adjusted at each time step. The coefficient values are, therefore, not fixed as in classical linear regression; they vary from year to year and could thus account for changes in yield trends (e.g., stagnation, increase or decrease in yield increase rate). This DLM approach has not yet been applied to analyses of yield time series. The aim of this study was to compare the performance of eight statistical models, including DLM, for analyzing yield time series and predicting yield trends. We chose wheat as the crop for this analysis, because wheat is an important cereal (27% of total cereal production worldwide) and because wheat yield time series show a great diversity of trends (increasing, plateauing, decreasing). We used a large number of wheat time series obtained at the national scale (in 120 countries) to compare the statistical models.

2. Material and method

The dataset includes wheat yield time series extracted from the FAOSTAT database for all countries, including data for wheat (120 countries). For most countries, yield time series began in 1961 and ended in 2010. The only exceptions were eight countries created after 1961, for which time series were shorter. In this dataset, bread wheat (*Triticum aestivum L.*) was not distinguished from durum wheat (*Triticum turgidum L.*) and winter wheat was not separated from spring wheat, because no distinction between these crops was made in the FAOSTAT database.
We assessed the suitability of seven different statistical models for analyzing yield time series. The models were first fitted to the time series included in our datasets and their qualities of fit were compared. The accuracy of the yield predictions obtained with the models was then assessed by cross-validation. The models considered in this study were: linear, quadratic, cubic, Holt-Winters

models (two variants; one with trend, one without trend), and two different dynamic linear models (noted DLMs and DLM0).
With DLMs, yield predictions were derived from the values of a_t and b_t were calculated with the Kalman smoother algorithm. The parameters a_t and b_t are defined as dynamic random variables, the values of which are estimated each time a new measurement becomes available. Their values are estimated by the conditional expected values of a_t and b_t given the available yield data, i.e., $E\left(a_t|Y_1,...,Y_M\right)$ and $E\left(b_t|Y_1,...,Y_M\right)$ where Y_1, ..., Y_M are the M yield data of the time series. The Kalman smoother algorithm can also be used to calculate the variances of the conditional probability distributions of a_t and b_t. The expected values and variances are calculated analytically with two equations; an observation equation relating yield data to the parameters a_t and b_t, and a system equation describing the changes in a_t and b_t from year to year. The observation equation is defined by

$$Y_t = a_t + b_t \times \Delta t + \varepsilon_t$$

with Δt =time since the last measurement, and $\varepsilon_t \sim N\left(0,\sigma_\varepsilon^2\right)$. Here, we will use Δt =1 year, because we will assume that a yield measurement is available every year. The system equation is defined by

$$Z_t = Z_{t-1} + \eta_{t-1}$$

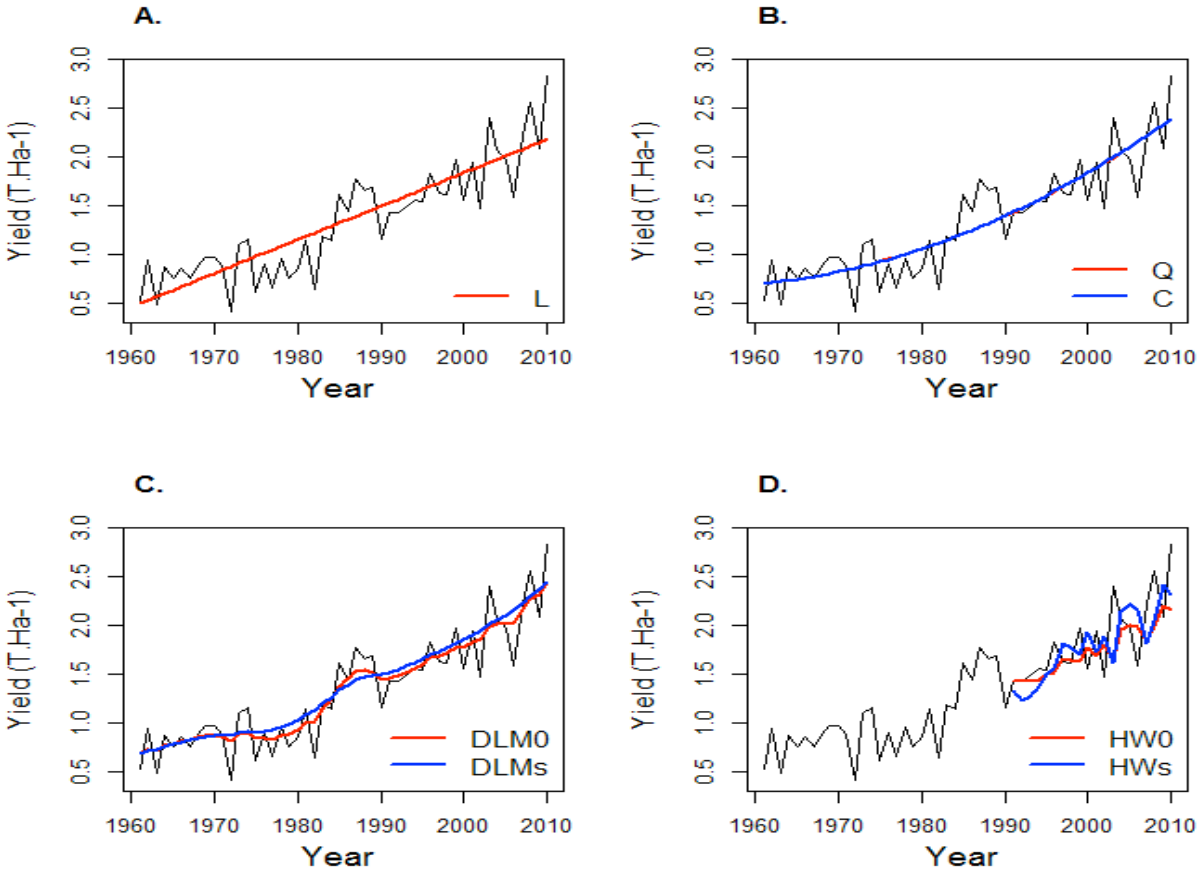
with $Z_t = \begin{pmatrix} a_t \\ b_t \end{pmatrix}$, $\eta_{t-1} \sim N\left(0,\Sigma\right)$, and $\Sigma = \begin{pmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_b^2 \end{pmatrix}$

In the DLMs model defined above, the state variables are the intercept and slope of a linear regression equation. They are assumed to vary from year to year according to a stochastic process defined by the system equation. The slope b_t corresponds to the yearly yield increase rate (i.e., the yield increase obtained in one year). The observation equation relates the yield data Y_t, t=1, ..., N, to the two state variables $Z_t = \begin{pmatrix} a_t \\ b_t \end{pmatrix}$, and the system equation relates the values of the two state variables at time t to the values at time t-1. The model includes three unknown parameters: $\sigma_\varepsilon^2, \sigma_a^2$, and σ_b^2. The variance σ_ε^2 quantifies the variability of yield around the trend. The variances σ_a^2 and σ_b^2 quantify the variability of the intercept and slope of the linear trend and define their change over time. The three variances σ_ε^2, σ_a^2, and σ_b^2 must be estimated from data. The DLM0 model is a simplified version of DLMs with b_t=0 and σ_b^2=0. DLM0 corresponds to a random walk model and includes only two parameters that need to be estimated from data.
Model parameters were estimated for each time series and for each country. The parameters of models L, Q, and C were estimated by ordinary least squares, using the function lm of R software. The model residuals obtained at different dates were not correlated, according to the autocorrelations calculated with the acf function of R. The parameters of the HW0 and HWs models were estimated with the optimizer of the HoltWinters function of R. The parameters of DLM0 and DLMs were estimated by maximum likelihood, with the function dlmMLE of the package dlm of R (Petris, 2010). Yield prediction accuracy was assessed by one-year-ahead cross-validation.

3. Results - Discussion

An example of wheat yield time series obtained in Brazil is shown in Figures 1. Yield was 0.5 t ha^{-1} in 1961, gradually increasing thereafter to reach almost 3 t ha^{-1} in 2010. The L model tended to overestimate yield between 1975 and 1985, and to underestimate yield between 1961 and 1972 (Fig. 1A). Visually, the yield trends fitted by Q and C were almost indistinguishable (Fig. 1B). With both models, the rate of yield increase tended to increase with time, especially after 1990. The yield trends obtained by fitting DLM0 and DLMs were very similar, but that obtained with DLMs was smoother (Fig. 1C). HWs predictions were almost always higher than those obtained with HW0 (Fig. 1D), due to the high rate of yearly yield increase observed in Brazil after 1990. This increasing trend was taken into account by HWs but not by HW0.

Figure 1: *Wheat yield time series in Brazil and fitted values obtained with different models. A: linear (L) models. B: Quadratic (Q) and cubic (C) models. C: dynamic linear models with and without trend (DLMs, DLM0). D: Holt-Winters model with and without trend (HWs, HW0).*



HW0 and DLM0 had the lowest RMSEP and gave the most accurate future yield predictions (Table 1). As the RMSEP values obtained with the HW0 and DLM0 models were very similar, it is difficult to choose between these two models for the prediction of future yields. The DLMs model was ranked third and the model with the least accurate predictions was the linear model L (Table 1).

Although DLMs did not lead to the most accurate predictions, this model has an interesting practical advantage, in that it can be used to estimate both yield levels and yearly yield increase/decrease rates (noted a_t and b_t above), whereas DLM0 and HW0 estimate only yield levels (i.e., only a_t). The estimation of yield increase rates is useful, because the values obtained

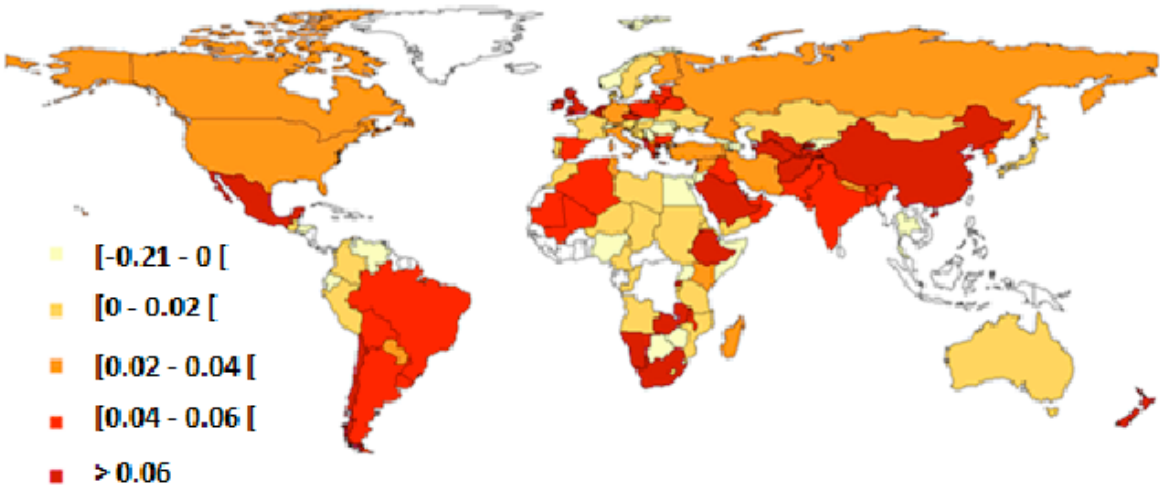
indicate whether yield is stagnating, decreasing or increasing in the geographical areas of interest. When estimated dynamically every year, yield increase rates reveal changes in yield trends over time and provide useful information about trend changes. Yearly yield increase/decrease rate is an important parameter is foresight studies on food security, because it is used to determine whether food and feed supplies fit food and feed demands. According to Ye et al. (2013), yield increase rate is a good indicator of food security. It is, therefore, useful to estimate this parameter from yield time series.

Table 1: *Root mean square error of one-year ahead predictions (RMSEP) obtained for several statistical models: linear regression (L), quadratic regression (Q), cubic regression (C), dynamic linear models with and without trend (DLMs, DLM0), and Holt-Winters with and without trend (HWs, HW0). The differences with respect to the lowest RMSEP values are expressed as a percentage.*

	Model						
Units	L	Q	C	DLMs	DLM0	HWs	HW0
t ha^{-1}	0.52	0.48	0.47	0.43	0.42	0.44	0.42
%	24.45	14.24	12.54	3.43	0.02	5.55	0.00

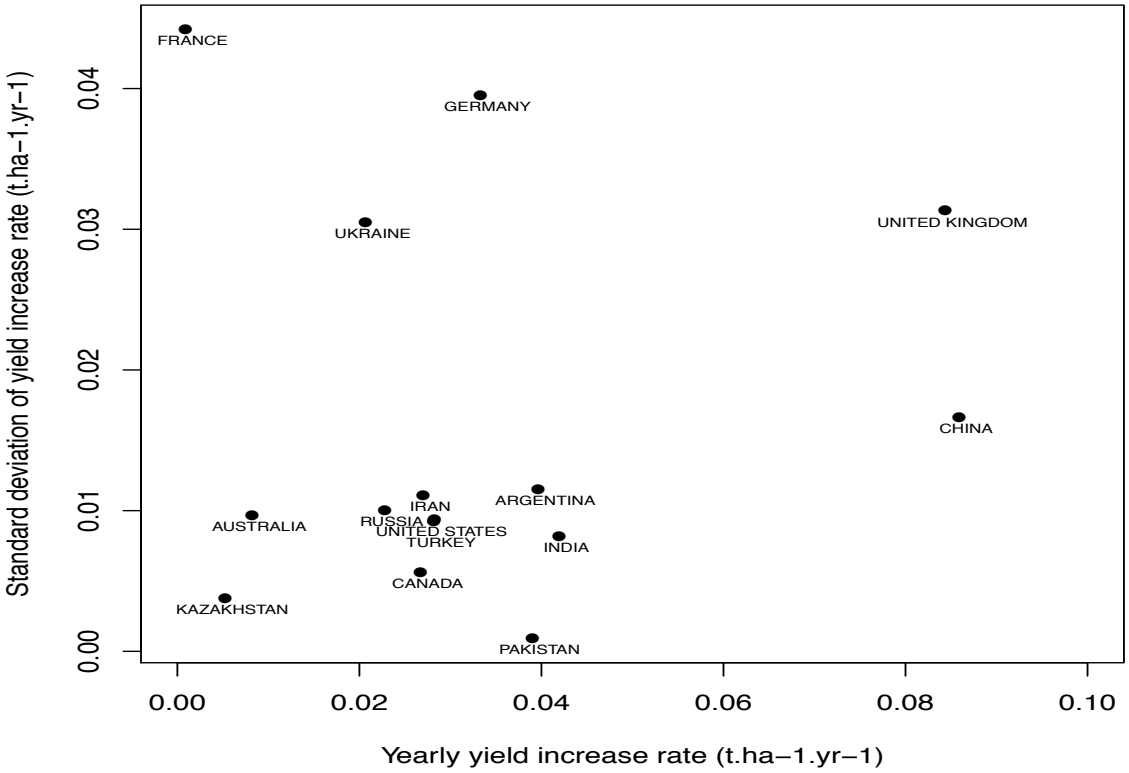
The practical value of DLMs is illustrated in Figure 2. The map in Figure 2 shows wheat yearly yield increase rates estimated for the world in 2010. These increase rates correspond to the conditional expected value of b_t obtained with DLMs. Its between-country variability was very high (Fig. 2). Wheat yields were found to be stagnating or even declining in many countries (in France, but also in Norway, Sweden, Portugal, and several countries in Eastern Europe, Africa and South America), but estimated yield increase rates were above 0.06 t ha^{-1} $year^{-1}$ in several countries in Europe, Asia, Africa and America.

Figure 2: *Estimated yield increase rates (t ha^{-1} $year^{-1}$) obtained with DLMs for wheat. Countries for which wheat yield data are not available are indicated in white.*



Another advantage of DLMs is that it can be used to calculate the conditional variances (or standard deviations) of the yield trends (i.e., the conditional variance of b_t) for the different countries. These variances indicate whether the yield trend is precisely estimated or not. Variances of yield trend estimated with DLMs are shown in Figure 3 for 15 countries. Some countries (e.g., France, Germany, UK) are characterized by high standard deviation showing that yield trend is not precisely estimated and may change upward or downward in the future (Fig. 3). For other countries such as Pakistan, Kazakhstan and Canada, standard deviations were close to zero indicating reliable estimated yield trends.

Figure 3: *Estimated yield increase rates and standard deviation of increase rate (t ha^{-1} year^{-1}) obtained with DLMs for wheat and for the year 2010. The 15 countries with the highest wheat productions in 2010 are presented on the graph.*



4. Conclusion

The Holt-Winters and dynamic linear models performed equally well, giving the most accurate predictions of wheat yield. However, dynamic linear models have two advantages over Holt-Winters models: they can be used to reconstruct past yield trends retrospectively and to analyze uncertainty. The results obtained with dynamic linear models indicated a stagnation of wheat yields in many countries, but the rate of increase of wheat yield remained above 0.06 t ha^{-1} year^{-1} in several countries in Europe, Asia, Africa and America.

REFERENCES

Brisson N., Gate P., Gouache D., Charmet G., Oury F-X., et al. (2010) Why are wheat yields stagnating in Europe? A comprehensive data analysis for France, *Field Crops Research*, 119, 201–212.

FAOSTAT: http://faostat.fao.org/.

Finger R. (2010) Evidence of slowing yield growth–the example of Swiss cereal yields, *Food Policy*, 35, 175–182.

Kumar-Boken V. (2000) Forecasting spring wheat yield using time series analysis, A*gronomy Journal*, 92, 1047–1053.

Petris G., Campagnoli P., Petrone S. (2009) Dynamic Linear Models with R, Springer, 258 p.

Petris G. (2010) An R package for dynamic linear models, *Journal of Statistical Software*, 36, 1–16.

Ye L., Xiong W., Li Z., Yang P., Wu W., Yang G., Fu Y., Zou J., Chen Z., Van Ranst E., Tang H. (2013) Climate change impact on China food security in 2050, *Agronomy for Systainable Development*, 33, 363-374.