HARVEST PREDICTION MODEL BASED ON PUBLIC DATA FOR LARGE REGIONS

Andrius Zuoza¹, Aurelijus Kazys Zuoza², Audrius Gargasas³

¹ Agro Baltic GmbH. Rungestr. 17, 18055, Rostock, Germany.
Tel. +491721987076. E-mail a.zuoza@agro-baltic.de
² Assoc. prof. dr. Lithuanian Sport University, ³ Prof. dr. Aleksandras Stulginskis university

Received 02 02 2015; accepted 03 03 2015

This article describe harvest prediction model for the country or for the big region on the public available data. In the article are analysed impact of main fertilizers component and environmental variables to the grain harvest. The aim of the article was to create regression model, which best describes grain harvest prediction on public (free) available data. Created final regression model explain 78% ($R^2$) of the variation in the harvest result. Presented model show, that prediction accuracy significantly increase if environmental variables are added. Prediction accuracy (RMSE) of the final regression model was 3.89. All calculation was made on the example of the Germany.

Keywords: agriculture; harvest; prediction; fertilizer, regression model.
JEL codes: C53, Q11.

1. Introduction

Prediction of the upcoming harvest is important not only for farmers, but also for local authority (Tenkorang, 2008), world aid organizations, food industry, transport companies, traders and others. The prediction model should be simple to calculate (Eckert, 2001), the data for the model – relative cheap, and accuracy of the prediction high at least one month before the starting of the harvest.

Depending on the available data and end-user purpose it can choose a couple of the methods for harvest prediction. Prediction models can be classified (Weissteiner, 2003) as follows.

Deterministic models. These models are based on physical, chemical and physiological process of the plants evolution. These models are often used for small areas. The obtaining of the data is complicated and expensive. Empiric-statistic models. These models most of all are based on collected statistical environmental data. The data is relative cheap or free, but they are published with big delay. Spectrometereological models. Mostly this is a mixture of deterministic and empiric-statistic models. In these models are used remote sensing data, aerial photo graph and historical statistical data. These models are good for small region or Parcel. The data are easy to buy, but it’s expensive.
Regression models have been criticized, since underlying mechanisms which transform climatic input into yield are not explicitly described and the hierarchical structure of the underlying physiological processes is not taken into account. Monthly climatic effects predicted by a regression model are not always easily interpreted from a physiological background because the model can only be an approximation of the underlying processes, and may fail to include some of them. Because of their empirical nature, regression models are restricted to the range of climate data from which they are developed (Landau, 2000).

There are large amount of the trails to create yield regression model based on climate change, but most of the model are based only on one subject, i.e. mostly only on environment data (such as temperature, humidity, etc.) (Lobell, 2010; R Core…, 2012) or on satellite image data (Liang, 2007; Weissteiner, 2003).

**Research object:** Grain harvest, fertilizer consumption, environmental variables and their influence to harvest prediction.

**Research method:** The following research methods were used: analysis of literature and document, statistical analysis of the data and standard multivariate linear regression analysis. Research data was analysed and regression models was make with programming language and software environment for statistical computing R (version 3.1.2).

The **aim** of the article is to create and present harvest prediction model on the public available data with high prediction rate at least one month before harvest.

Data for this prediction model was collected from public sources. These data were collected for the calculations:

- grain harvest in Germany from 1994 till 2013 in million metric tons; (Wodsak, 2013);
- fertilizer consumption by main components (Nitrogen, P_{2}O_{5}, K_{2}O, CaO in thousand tons of active material) in Germany from 1993/1994 season till 2012/2013 season, (Wodsak, 2013);
- average air temperature (in °C), average precipitation (in mm) and sunshine duration (in hours) from 1993 till 2013. (Germany's, 2014).

The following manipulation was made with the original data:
- fertilizer season starts at 1st of the July and ends at 31st of June;
- value of the grain harvest per calendar year was assigned to the fertilizer season. E.g. if the harvest was in 1994, then it will be assigned to 1993/1994 season;
- average air temperature, precipitation and sunshine duration was recalculated from calendar year to the fertilizer season;
- average air temperature, precipitation and sunshine duration was split to 1st half of the season and to the 2nd half. July-December was assign to the 1st half of the season and January-June was assign to the 2nd half of the season.

All data manipulation was made with R programming language (R Core…, 2012).
2. Exploratory analysis

Exploratory analysis was performed by examining tables and plots of the observed data. There were identified transformations to perform on the raw data on the basis of plots and knowledge of the scale of measured variables.

First was checked does harvest data fit normal distribution. The result of the Shapiro–Wilk normality test proved that the data are normally distributed (W = 0.97, p > 0.05).

Based on publication (Agada, 2006) it was expected, that nitrogen consumption will make biggest influence to the grain harvest.

Correlation analysis of the data shown, that there are no strong correlation between harvest and all other data.

The strongest and statistical significant correlation was found between harvest and P<sub>2</sub>O<sub>5</sub> fertilizers (r = −0.46; p < 0.05) and CaO fertilizers (r = 0.53; p < 0.05). All others correlation coefficients associated with harvest wasn’t statistical significant.

To relate harvest to gather data were performed a standard multivariate linear regression model (Seber, 2012). Model selection was performed on the AIC in a stepwise algorithm (R Core…, 2012).

3. Results

The harvest data used in this analysis contains information about grain harvest (H), nitrogen (N), phosphor (P<sub>2</sub>O<sub>5</sub>), potash (K<sub>2</sub>O), calcium (CaO) fertilizer consumption (in thousand metric tons), yearly sunshine duration (Sun), sunshine duration in first half (Sun1H) and second half (Sun2H) of season, average temperature of the season (Temp), average temperature in first half (Temp1H) and second half (Temp2H) of season, average precipitation of the season (Pre), average precipitation in first half (Pre1H) and second half (Pre2H) of season.

There were identified no missing values in the collected data set and all measured variables were observed to be inside the standard ranges. Harvest data in this data set also did not seems to show any patterns over the time.

To test regression model and to calculate true prediction error data were split to train and test data. Train data – all data from 1993/94 season till 2009/10 season and test data – from 2010/11 season till 2012/13 season. Harvest data are normally distributed and 95 percent confidence interval is 41.9–46.0.

There were made two regression models: one only with the consumption of the main fertilizer components and second with consumption of the main fertilizer components and environmental factors. To compare regression models, there were compared adjusted R<sup>2</sup> and root mean square error (RMSE). RMSE is one of the most commonly used statistical parameters, which represents the average difference between estimated and observed value.

The RMSE was calculated as a better evaluation method, but can also it can explain whether the model under- or over-predicted (Liang, 2007). The adjusted R<sup>2</sup> score of the regression model which explore only main fertilizer components (N, P<sub>2</sub>O<sub>5</sub>, K<sub>2</sub>O) influence to harvest was 0.46. Model is shown by this equation (2):
\[ H = b_0 + b_1 \times N + b_2 \times P_2O_5 + b_3 \times K_2O \]  

(2)

Coefficients and p-values of this model are shown in Table 1.

<table>
<thead>
<tr>
<th>Estimate</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>b_0</td>
<td>14.925</td>
</tr>
<tr>
<td>b_1</td>
<td>0.026</td>
</tr>
<tr>
<td>b_2</td>
<td>-0.07</td>
</tr>
<tr>
<td>b_3</td>
<td>0.011</td>
</tr>
</tbody>
</table>

Table 1. Regression models coefficients

To make more précised model there were used Akaike information criterion (AIC). The model was constructed with all possible inputs from database available and then selected the best fitted model. It became that the best model have the following equation (3):

\[ H = b_0 + b_1 \times N + b_2 \times P_2O_5 + b_3 \times Pre1H + b_4 \times Temp + b_5 \times Temp1H + b_6 \times Sun \]  

(3)

The adjusted R^2 score of this (3) regression model is 0.78. Coefficients and p-values of this model are shown in Table 2.

<table>
<thead>
<tr>
<th>Estimate</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>b_0</td>
<td>69.709</td>
</tr>
<tr>
<td>b_1</td>
<td>0.027</td>
</tr>
<tr>
<td>b_2</td>
<td>-0.062</td>
</tr>
<tr>
<td>b_3</td>
<td>-0.038</td>
</tr>
<tr>
<td>b_4</td>
<td>1.279</td>
</tr>
<tr>
<td>b_5</td>
<td>-3.189</td>
</tr>
<tr>
<td>b_6</td>
<td>-0.014</td>
</tr>
</tbody>
</table>

Table 2. Regression models coefficients

Our regression model shows, that if nitrogen consumption increased by one thousand metric tons (i.e. one unit) then grain harvest will increase by 27 thousand metric tons, holding that all other predictors are fixed. The regression model shows, that the opposite effect occurs if the consumption of the \( P_2O_5 \) will increase by one unit (i.e. one thousand metric tons of \( P_2O_3 \)), then grain harvest will decrease by 62 thousand metric tons, holding that all other predictors are fixed. The interpretation of temperature component in our regression model shows that average year temperature increased by one unit (i.e. 1 °C), then grain harvest will increase by abt. 1200 thousand metric tons, holding that all other predictors are fixed. But if the average temperature for the first half of the year will increased by one unit (i.e. 1 °C), then grain harvest will decrease by abt. 3800 thousand metric tons, holding that all other predictors are fixed.

Predicted values from the both models are shown in Table 3.
### Table 3. Predicted values by regression models

<table>
<thead>
<tr>
<th>Season</th>
<th>Observed value</th>
<th>Predicted value by model 1</th>
<th>Predicted value by model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010/2011</td>
<td>41.1</td>
<td>47.14</td>
<td>43.91</td>
</tr>
<tr>
<td>2011/2012</td>
<td>45.4</td>
<td>45.45</td>
<td>44.09</td>
</tr>
<tr>
<td>2012/2013</td>
<td>47.1</td>
<td>43.50</td>
<td>44.76</td>
</tr>
</tbody>
</table>

Calculated root mean square error of the model 1 was 7.03 and of the model 2 – 3.89.

### 4. Conclusions

1. The presented analysis shows that it is possible to predict oncoming harvest with data from public (free) sources. It also shows that harvest prediction is more precise when the environmental variables are added. This analysis estimates positive relationship using linear model relating harvest and average temperature and negative relationship between harvest and temperature of the 1st half of the season and sun radiation.

2. Presented models shows, that prediction accuracy significantly increase if environmental variables are added. It is possible to make more précised prediction model by adding more variables e.g. harvest prices for season or fertilizer consumption and environmental variables for separate regions of the country, or to investigate influence of environmental variables in different time periods (for example – monthly based).

3. By author’s opinion, expanded and more precise model could be interesting for a lot of business and governmental organizations.

### Literature


- Germany’s National Meteorological Service (Deutscher Wetterdienst). Climate data for German areas and regions. – http://www.dwd.de/bvbw/appmanager/bvbw/dwdwwwDesktop?_nfpb=true&_pageLabel=_dwdwww_menu2_leistungen_a-z_freiemetinfos&T115202758871200642573928gsbDocumentPath=Navigation%2FOeffentlichkeit%2FKlima__Umwelt%2FKlimadaten%2FKldaten__kostenfrei%2Fdaten__gebietsmittel__node.html1%3F__nnn%3Dtrue[01 08 2014].


DERLIAUS PROGNOZĖS DIDELEMS REGIONAMS MODELIS PANAUDOJANT VIE-ŠUOSIUS DUOMENIS

Andrius Zuoka1, Aurelijus Kazys Zuoka2, Audrius Gargasas3
1 Agro Baltic GmbH. Rostock, Germany, 2 Lithuanian Sport University, 3 Aleksandras Stulginskis university

Įteikta 2015 02 02; priimta 2015 03 03

Santrauka


Raktiniai žodžiai: žemės ūkis, derlius, prognozė, trąšos, regresijos modelis.
JEL kodai: C33, Q11.