The CGMS Statistical Tool for yield forecasting

Steven Hoek, Paul W. Goedhart and Wies Akkermans

Wageningen UR, P.O. Box 9101, 6700 HB Wageningen, the Netherlands. Tel. +31-(0)317-481718. Email: steven.hoek@wur.nl

Introduction

Official EU forecasts for crop yields are calculated several times per year by crop analysts of the Joint Research Centre (JRC). The analysts determine the forecasts by means of the MARS Crop Yield Forecasting System (MCYFS). In general the forecasts for a certain region are based on statistical models which describe historical yields in terms of a time trend combined with a relationship with CGMS indicator data. For a general overview of the CGMS, see the article by Boogaard et al. (this issue).

So far, it is mostly simulated crop yields which are used as indicator data. Attempts are underway to also include (1) weather indicator data and (2) remote sensing data. In doing so, it may be possible to get around the limitations of the used crop simulation models. Ideally, the influence of weather indicators - as experienced in the course of the growing season - are reflected in the simulated crop yields. However, effects of extreme weather conditions at specific points in time – e.g. during flowering or harvesting – are often not reflected. The accumulated effects of weather on vegetation and crops are sometimes better reflected by remote sensing indices such as vegetation indices than by simulated crop yield data.

Nevertheless, expectations should not become too high. It is unfeasible to include all possible environmental factors into the statistical models. It should be realised that other factors of socio-economic and technological nature also influence the crop yields. The insights of the crop analysts are therefore expected to remain important in the process of obtaining reliable yield forecasts.

The CGMS Statistical Tool

Since 1994 the so-called CGMS statistical module has been in use at JRC - to facilitate crop yield forecasting at national and sub-national level. Now, an improved version has been developed, for use by the crop analysts of JRC. The tool was developed by two subdivisions of the Wageningen UR: Alterra-CGI and Biometris. Aim was to enable the analysts to construct more elaborate models (regression and scenario models) than they could construct with the previous module and to make it easier to include more indicators, i.e. weather indicator and remote sensing data.

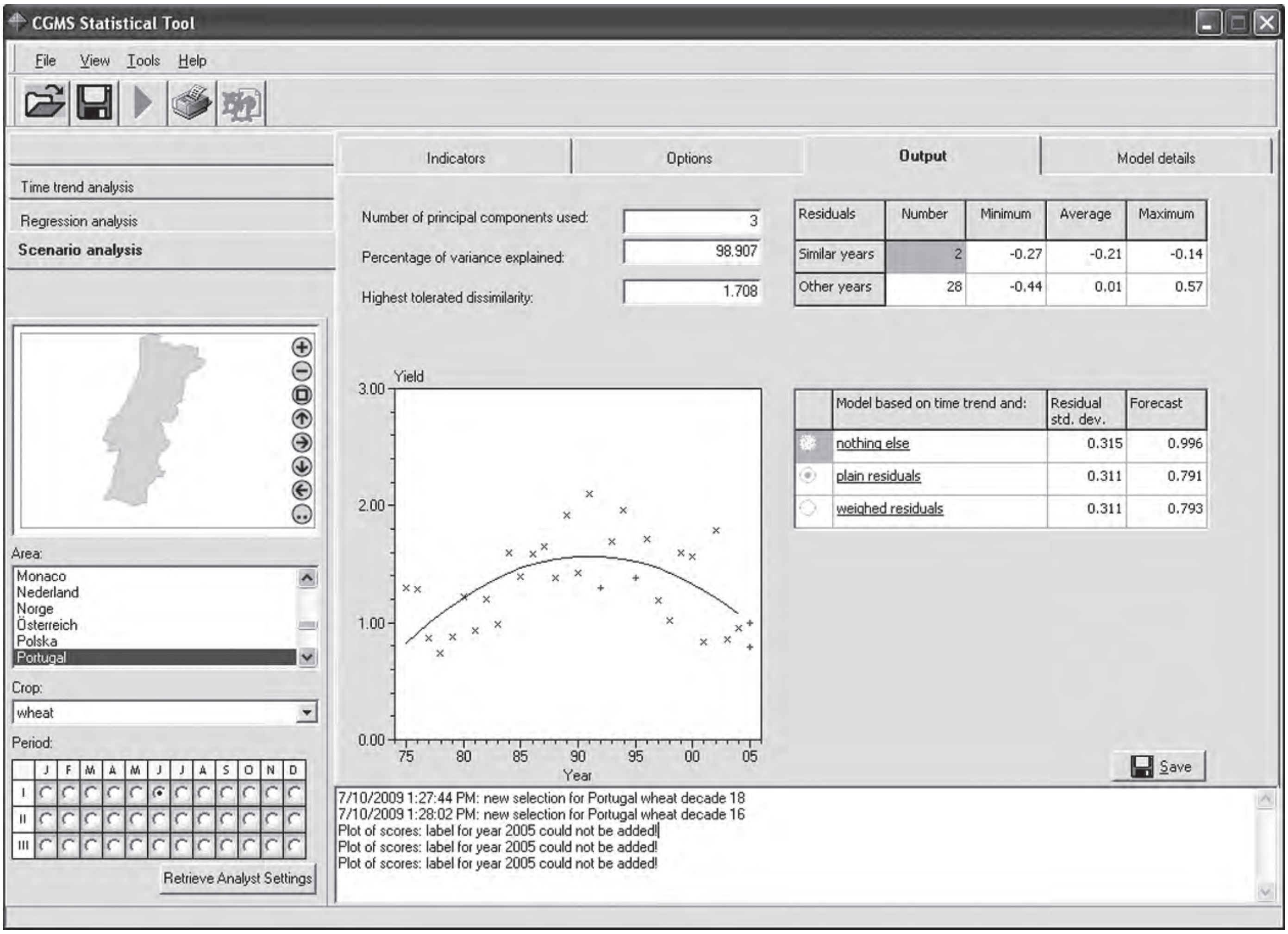
Other features of the new CGMS statistical tool:

- Data retrieved directly from the CGMS database
- The analyst is guided through the process by a series of steps
- Visualisation of the yield data and presentation of other data on the screen
- Improved support for comparison and selection of models
- Model settings for future use as well as for sharing with colleagues
- An interactive mode as well as a batch mode; in the latter mode, models are constructed as guided by model settings saved or compiled earlier on
- Report feature, enabling the user to easily present a constructed model to others.

Statistical aspects

The regression and scenario models have in common that they include a mean regional yield as well as a time trend of regional historical yields. Time trend analysis is therefore always required. The time trend represents the influence of long-term economic and technological dynamics such as increased fertilizer application, improved crop management, improved varieties etc. on the regional yields and production. The time trend can be modelled by means of a linear or quadratic function of time. Graphs and statistical test are provided to guide in the selection of the appropriate time trend. If necessary, a logarithmic transformation can be performed first. In practice, the time trend never explains all the variation in the yields. Regression and scenario analysis basically are two different ways for describing the remaining variation.

In the case of regression analysis, the remaining variation is modelled as a function of indicators. Many indicators may be presented, but in general only a few are relevant. The analyst is guided to select the best model; statistical tests and summary regression statistics are calculated, such as R-squared. So far only crop growth simulation results have been used. These simulation results represent the inter-annual yield variation that results from weather variability. For instance the vector of simulated biomass of June 10 of the last 10 previous years is included into a regression model in order to explain the vector of historical yields for the same 10 years. After the model has been determined, the simulated biomass for June 10 for the current year can be entered into the



Number of principal components used: 3

Percentage of variance explained: 98.907

Highest tolerated dissimilarity: 1.708

Residuals	Number	Minimum	Average	Maximum
Similar years	2	-0.27	-0.21	-0.14
Other years	28	-0.44	0.01	0.57

	Model based on time trend and:	Residual std. dev.	Forecast
○	nothing else	0.315	0.996
●	plain residuals	0.311	0.791
○	weighed residuals	0.311	0.793

Area:
Monaco
Nederland
Norge
Österreich
Polska
Portugal

Crop: wheat

Period:

	J	F	M	A	M	J	J	A	S	O	N	D
I	○	○	○	○	○	●	○	○	○	○	○	○
II	○	○	○	○	○	○	○	○	○	○	○	○
III	○	○	○	○	○	○	○	○	○	○	○	○

```
7/10/2009 1:27:44 PM: new selection for Portugal wheat decade 18
7/10/2009 1:28:02 PM: new selection for Portugal wheat decade 16
Plot of scores: label for year 2005 could not be added!
Plot of scores: label for year 2005 could not be added!
Plot of scores: label for year 2005 could not be added!
```

Figure 1 The CGMS Statistical Tool showing 3 scenario models with a quadratic time trend pertaining to the wheat yield in Portugal (in red the forecasted yield range based on the 2 similar 1992 and 1995 after corrected for the quadratic time trend).

model equation as soon as it becomes available, leading to the predicted final yield for the current year.

In the case of scenario analysis, the indicator values for the current year are required for the model construction. The Euclidean space formed by the indicator vectors of all the years – incl. the current one - is analysed first by means of Principal Component Analysis, so that a model is obtained which explains e.g. 95% of all the variation found in the indicator vectors; subsequently the years are categorised into (1) years which are similar to the current one and (2) other years. The forecast for the current year is then calculated on the basis of the yields which were realised in the similar years. Figure 1 shows the result of a scenario analysis done with the tool.

Software engineering aspects

The tool was developed in Delphi 7 but the regression routines were implemented in FORTRAN using the IMSL library and compiled into a dynamic link library (DLL). Various third-party Delphi component packages were used incl. SDL MathPack & ChartPack from Epina.

Already in the previous module, database access was arranged by means of the Borland Database Engine (BDE) and this was left unchanged. The used SQL queries were stored in text files and can be modified to accommodate limited changes to the database structure. For the previous version a simple object-relational mapping framework was used - developed in-house - and this was reused. The tool can be configured to work with an Access database or with an Oracle database. The tool only runs on Windows.