The CGMS Statistical Tool

User Manual

Paul W. Goedhart[1] & Steven B. Hoek[2]

ASEMARS Project Report no. 1.4.1

September 2008

[1] *Biometris, Wageningen UR, the Netherlands*
[2] *Alterra, Wageningen UR, the Netherlands*

Commissioned by Joint Research Centre of the European Commission, Ispra, Italy

The CGMS Statistical Tool

User Manual

Paul W. Goedhart and Steven B. Hoek

Wageningen-UR, Wageningen, 2008

ABSTRACT

Paul W. Goedhart & Steven B. Hoek, 2006. The CGMS Statistical Tool. User Manual. Wageningen, Wageningen-UR, 77 pp.

The CGMS statistical tool has been developed for JRC's MARS project in the framework of the contract study "Actions in Support of the Enlargement of the MARS Crop Yield Forecasting System (MCYFS) Lot I (ASEMARS Lot I)". The tool is designed for use by the crop analysts and is an improved version of the CGMS statistical module which was in use since 1994 to facilitate national and sub national crop yield forecasting. Time trend analyis of yield statistics is followed by regression or scenario analysis using biophysical indicators to explain yield statistics and search for similar years. Constructed models are used to predict yield of the current growing season.

Keywords: CGMS, MCYFS, crop yield forecast, crop yield statistics

Contents

Preface

The CGMS statistical tool has been developed for JRC's MARS project in the framework of the contract study "Actions in Support of the Enlargement of the MARS Crop Yield Forecasting System (MCYFS) Lot I (ASEMARS Lot I)" by two institutes of the Wageningen University and Research Centre: Alterra and Plant Research International. The tool is designed for use by the crop analysts of the MARS project at JRC and is an improved version of the CGMS statistical module which was in use since 1994 to facilitate national and sub national crop yield forecasting.

The authors would like to thank Giampiero Genovese, Manola Bettio, Bettina Baruth and Iacopo Cerrani of the MARS STAT team of the Joint Research Centre for very fruitful discussions; it was a pleasure to visit you in Ispra. We are indebted to Yannick Curnel and Roger Oger of the Walloon Agricultural Research Centre for a very thorough and detailed validation of a beta version of CgmsStatTool. We would also like to thank Hendrik Boogaard, Allard de Wit and Kees van Diepen of Alterra for their cooperation.

Summary

The CGMS statistical tool has been developed for JRC's MARS project in the framework of the contract study "Actions in Support of the Enlargement of the MARS Crop Yield Forecasting System (MCYFS) Lot I (ASEMARS Lot I)" by two institutes of the Wageningen University and Research Centre: Alterra and Plant Research International. The tool is designed for use by the crop analysts of the MARS project at JRC and is an improved version of the CGMS statistical module which was in use since 1994 to facilitate national and sub national crop yield forecasting.

The improvement of the statistical module of CGMS is part of a larger research effort launched by the Agrifish Unit to complete and reinforce the CGMS in order to extend the system thematically and geographically in support to the operational activities, and to improve the efficacy of the system, and flexibility. The improvement of the statistical module associated to CGMS involves the extension of the functionalities of the system based on the following requirements:

- Allow to select from the data base any meteo / crop / RS parameter individually as indicator (for a maximum of 20 predefined parameters) in the regression analysis;
- The implementation of automatic multiregressive approaches in order to let emerge the best model;
- Automatic performances check on the regression models obtained and statistical tests to take the decision on which is the best model that must be proposed as final each decade running;
- Complete the implementation of the scenario analysis (as currently applied by MARS-STAT) by adding the whole set of intermediate and final parameters available from CGMS, the satellite indicators from the MARSOP2 project.
- And implementing the algorithm for the prediction calculations.

This report describes the new statistical subsystem, which will be called CGMS Statistical Tool or CgmsStatTool for short. The interface has been rebuilt almost completely, again using Delphi. The underlying statistical functionality has also been revised completely. Fortran was still used as the programming language and high quality routines of the IMSL Fortran library are being invoked, e.g. for fitting a single regression model.

This report is not meant as an introduction to linear regression. It is therefore assumed that users of CgmsStatTool have a firm understanding of the basic principles of linear regression analysis. Before using the tool in an operational way, the user is strongly advised to play with the tool and to read the chapter "Some Statistical Issues". An excellent and complete introduction into linear regression is the book by Montgomery, Peck and Vining (2001). Chapters 2-8 of this report describe the CgmsStatTool interface and purpose in detail. Chapter 9 contains

important remarks on several statistical issues related to proper use of the tool. Chapters 10 and 11 describe technical aspects like the installation, the databases used by the program, the way in which analyst settings can be saved, copied, modified and shared, a description of the menu items and clickable icons and the batch mode. Chapter 12 provides some details about the validation of the tool, and references are listed in Chapter 13. A detailed description of the most important Fortran routines is given in the annexes, as well as details about the databases and user settings.

1 Introduction

1.1 The CGMS Statistical Tool in the context of the MARS Project

The CGMS statistical tool has been developed for JRC's MARS project in the framework of the contract study "Actions in Support of the Enlargement of the MARS Crop Yield Forecasting System (MCYFS) Lot I (ASEMARS Lot I)" by two institutes of the Wageningen University and Research Centre: Alterra and Plant Research International. The tool is designed for use by the crop analysts of the MARS project at JRC and is an improved version of the CGMS statistical module which was in use since 1994 to facilitate national and sub national crop yield forecasting.

The MARS project, for "Monitoring Agriculture with Remote Sensing" started in 1988 and was initially designed to provide to the DG Agriculture and DG EUROSTAT, independent and timely information on crop areas and yields. Since 1993 this information is brought together in the MARS Bulletin with early crop forecasts during the campaign till the obtaining of the official EU statistics. The interest of this activity is to provide every month, independently of Member States, real-time information on productions expected and to identify regional anomalies. MARS Bulletin is based on analysis of meteorological conditions, of results of crop growth simulation models using meteorological observations and high temporal resolution satellite data (VEGETATION or NOAA/AVHRR). It concerns European Union, Candidate Countries, Eastern European countries and Maghreb. Since 2001, the MARS Bulletin outputs are used as official crop forecasts by DG Agriculture and are integrated in the EUROSTAT crop forecast system. Also in 2001, the MARS project became an independent Unit and was reattached to the IPSC (Institute for the Protection and the Security of the Citizen). Since June 2004 the Unit became "Agriculture and Fisheries (AGRIFISH Unit)".

The analytical procedures start with screening and statistical analyses of data produced in the MCYFS and this forms the basis for crop yield forecasting at national and European level. The agro-meteorological model used within the MCYFS in MARS is the CGMS (Crop Growth Monitoring System). The core of the model for crop growth simulation consists in an adaptation of the WOFOST model (see http://www.alterra.nl) to the European context and as for Rye Grass consists in adaptation of the LINTUL model (Light Interception and Utilisation Simulator) to the European scale called LINGRA (LINTUL GRAssland).

The crop simulation model WOFOST calculates, among other model outputs, potential yield and potential total biomass as a function of weather and soil conditions. The WOFOST model integrates all the effects of the varying weather conditions which give different potential yields over the years. The model output is aggregated from single land units to administrative regions. In 1992 and 1993 it was studied whether regionally aggregated output of the WOFOST crop model could be

used for regional crop yield forecasting. This was done by regressing the yearly official yields onto the model output of WOFOST. Because the official yields frequently showed a yearly increase, a technological linear trend was added to the regression model if necessary. Since the fitted relations were adequate for prediction purposes, a statistical module for CGMS level 3 was developed. The module selected from four candidate WOFOST model outputs, further called indicators, the best performing indicator (e.g. potential total biomass or potential yield). The 'best' model, with a single indicator, was then used for forecasting the yield in the current year. In the operational CGMS system this was done for each crop region combination at the end of each decade (10 day period) during the agricultural season. This first version of the statistical subsystem employed a Fortran executable. The system was rebuilt in the year 2000 to enable the use of a Fortran DLL within S PLUS. The Fortran DLL that selects and calculates the regression models did not change. However, S Plus did not function very well in an operational production line. Therefore in CGMS version 8.0 the selection of data and indicators was organized in the user interface of C++. A dedicated Delphi program served as a statistical engine and called the Fortran DLL for performing linear regression.

1.2 The improvement of the statistical module of CGMS in ASEMARS

The improvement of the statistical module of CGMS is part of a larger research effort launched by the Agrifish Unit to complete and reinforce the CGMS in order to extend the system thematically and geographically in support to the operational activities, and to improve the efficacy of the system, and flexibility. The improvement of the statistical module associated to CGMS involves the extension of the functionalities of the system based on the following requirements:

- Allow to select from the data base any meteo / crop / RS parameter individually as indicator (for a maximum of 20 predefined parameters) in the regression analysis;
- The implementation of automatic multiregressive approaches in order to let emerge the best model;
- Automatic performances check on the regression models obtained and statistical tests to take the decision on which is the best model that must be proposed as final each decade running;
- Complete the implementation of the scenario analysis (as currently applied by MARS-STAT) by adding the whole set of intermediate and final parameters available from CGMS, the satellite indicators from the MARSOP2 project.
- And implementing the algorithm for the prediction calculations.

1.3 Guide for reading this Manual

This report describes the new statistical subsystem, which will be called CGMS Statistical Tool or CgmsStatTool for short. The interface has been rebuilt almost completely, again using Delphi. The underlying statistical functionality has also been revised completely. Fortran was still used as the programming language and high quality routines of the IMSL Fortran library are being invoked, e.g. for fitting a single regression model. This implies that future updates of the Statistical Module must be compiled with the Intel Visual Fortran Compiler Professional Edition which includes the IMSL library.

This report is not meant as an introduction to linear regression. It is therefore assumed that users of CgmsStatTool have a firm understanding of the basic principles of linear regression analysis. Before using the tool in an operational way, the user is strongly advised to play with the tool and to read the chapter "Some Statistical Issues". An excellent and complete introduction into linear regression is the book by Montgomery, Peck and Vining (2001). Users of CgmsStatTool are encouraged to read the following chapters of this book:

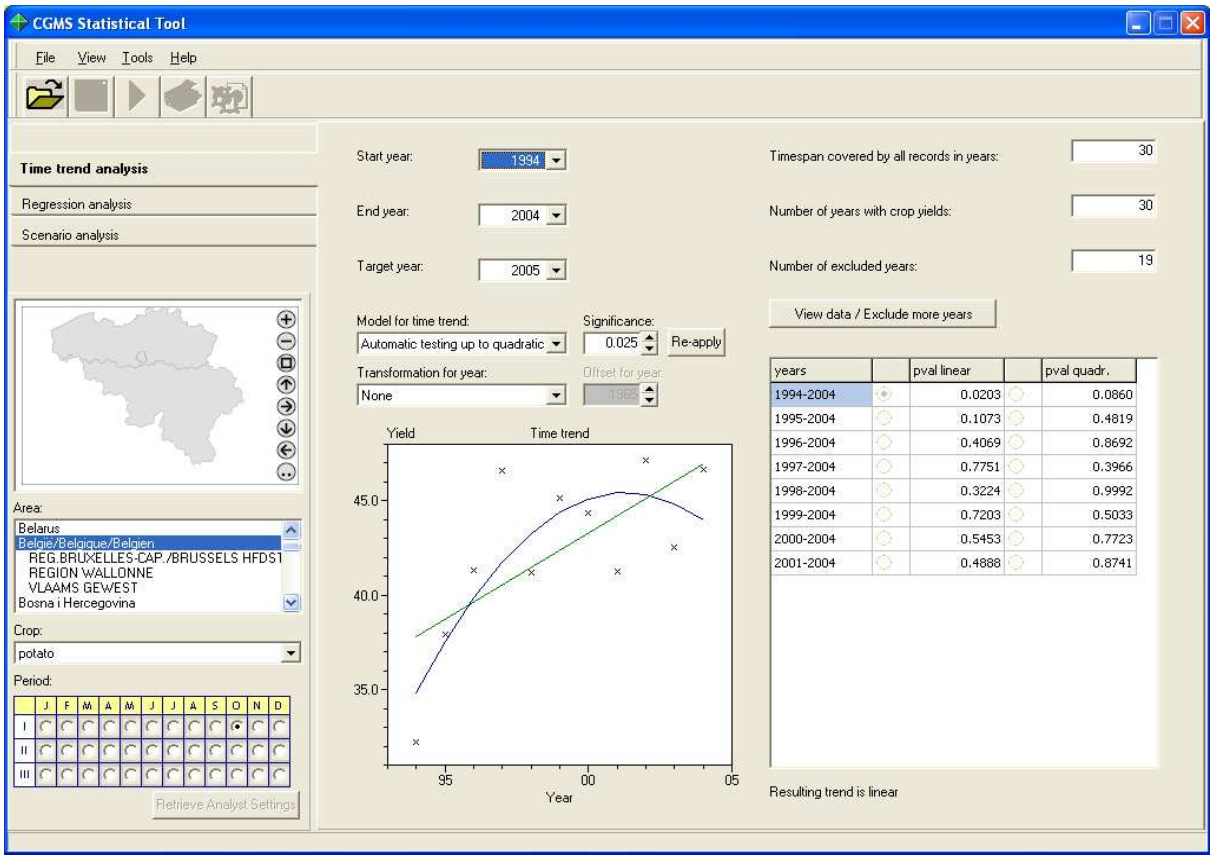
3. Multiple Linear Regression;
4. Model Adequacy Checking;
6. Diagnostics for Leverage and Influence;
9. Variable Selection and Model Building;
10. Multicollinearity.

Montgomery, Peck and Vining (2001) is frequently quoted, sometimes not explicitly, and referenced in this report. References are denoted by MPV followed by the relevant page number.

Chapters 2-8 of this report describe the CgmsStatTool interface and purpose in detail. Chapter 9 contains important remarks on several statistical issues related to proper use of the tool. Chapters 10 and 11 describe technical aspects like the installation, the databases used by the program, the way in which analyst settings can be saved, copied, modified and shared, a description of the menu items and clickable icons and the batch mode. Chapter 12 provides some details about the validation of the tool, and references are listed in Chapter 13. A detailed description of the most important Fortran routines is given in the annexes, as well as details about the databases and user settings.

2 Overview of Interface and Functionality

This chapter gives a brief overview of the interface and functionality of CgmsStatTool. A detailed description is given in subsequent chapters. Running CgmsStatTool in playground mode presents the user with the following opening screen.



The screen is divided into four parts:

1. A top row with a few menu items and a row with clickable icons for specific actions (read more in section 11);
2. The bottom left panel can be used to specify an area or region. This can either be done by clicking somewhere on the shown map, possibly after zooming, or by selecting an area from a drop down list. Once the Area is selected the available crops are shown in a second drop down list. After selecting a crop, the user must select a period (decade) for which a prediction must be made. By default this is the current period. (read more in Chapter 3);

3. The right panel consists of the following three tab pages:
   - A page for Time trend analysis; this page can be used to specify the calibration period, the target year for which the yield must be predicted, the time trend model and possibly a logarithmic transformation for year. An Excel type view of the data is available from this page (read more in Chapter 4);
   - A page for Regression analysis;
   - A page for Scenario analysis.
4. A fourth optional panel, at the bottom of the screen, can be shown by the {View | Log Window} menu item. This displays various warnings and errors that might occur, e.g. when indicators are aliased or when a perfect fit is obtained for a linear time trend model.

The user can open any of the three tab pages for analysis by using the vertical tabs which are placed on the left between the top row with clickable icons and the bottom left panel with controls for selecting area, crop and period. It should be noted that regression and scenario analysis are always based upon time trend analysis.

Both the tab page for regression analysis and the one for scenario analysis in turn consist of four tab pages which can be opened using horizontal tabs:

- The Indicators page can be used to select indicators which should enter the regression model or should be used in the scenario analysis. In case of regression analysis indicators can be either free or forced. Forced indicators are included in every regression model, while free indicators are either included or excluded from a model. The correlation matrix between the selected indicators can be viewed from this page (read more in Chapter 5)
- The Options page presents the user with all options. In case of regression analysis the main choice is between the single free indicators method and the method of best subset selection. The single free indicators method fits models with only one free indicator, in addition to the chosen time trend and forced indicators, if any. The best subset selection method fits the best models, according to some criterion, with multiple free indicators. There are various options to aid the user in selecting a proper model for prediction of the target year (read more in Chapter 6)
- The Output page displays the various, single indicator or best subset, regression models and results of the scenario analysis. Criteria for the different models are displayed as well as t values of included indicators. The t values can be coloured according to the sign or the significance of the corresponding regression coefficient. In case of scenario analysis only one criterion is presented: residual standard deviation (read more in Chapter 7)
- The Model details page is activated by clicking on a single model on the Output page. A detailed analysis of the single model is presented including a description of the model, summary statistics, regression

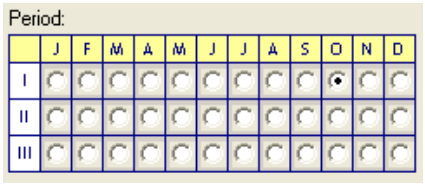
coefficients, case statistics such as fitted values, residuals and leverages, and finally a graphical representation of the case statistics (read more in Chapter 8).

3 Selecting an Area, Crop and Decade

A country can be selected by clicking on the name in the Area list box on the left or by clicking on the country shown as part of the map of Europe. When a country is selected, lower NUTS areas for that country then appear indented below the country name. A lower NUTS area can therefore only be selected after the relevant country has been selected first. The same is true for NUTS areas at lower levels.

Upon each selection in the left panel, a new query is sent to the database to update the time trend page on the right. If no data are available for the selected area, an informational message appears in the log window: "No crop statistics available for this NUTS area". The Crop list box will be updated to show the crops for which there are data available in the selected area.

Finally, the user has to select the decade for which the analysis should be carried out. The term decade refers to a ten day period. A month is considered to consist of three decades, the first taking from day 1 to day 10, the second from day 11 to day 20 and the last from day 21 to the end of the month. Decades are sometimes indicated by the name of the month followed by a Roman figure: I, II or III; at other times they are indicated by a number in the range 1 through 36. Selection is done by clicking one of the radio buttons of the so called decade selector:



The columns in this decade selector represent the months of the year. Selecting a different decade usually does not cause any change in the yield data shown on the time trend page. However, the indicator data are affected by the decade selection.

After selecting area, crop, decade and analysis type (regression or scenario), the user may press the button "Retrieve Analyst Settings" placed at the bottom of the left panel. When doing so, an attempt is made to retrieve all the settings for the selected Area, Crop, Decade and Analysis type combination from the analyst settings. However, if such settings were not saved to a so-called CSV file before, a message appears in the log window: "User defaults were not found - reverting to system defaults". Note that the function "Retrieve Analyst Settings" does not work for the analysis type Trend.

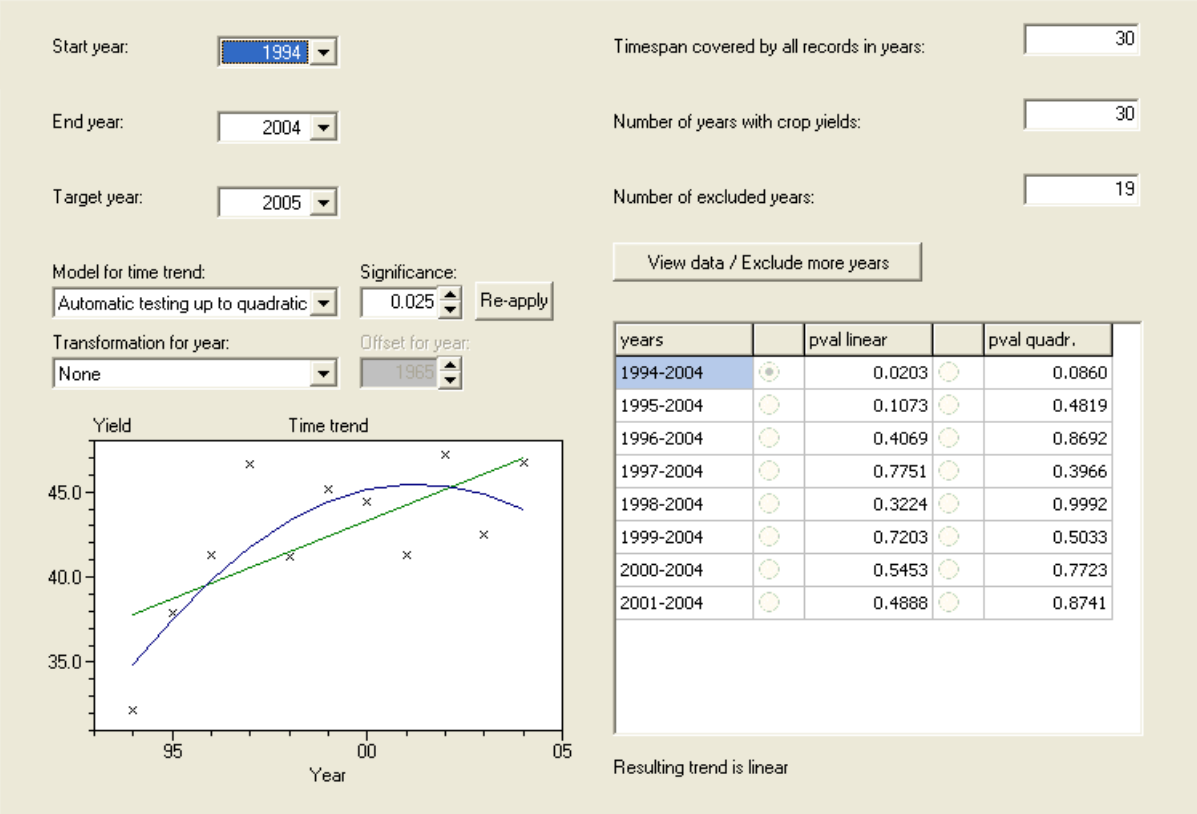
More information on analyst settings can be found in the section 10.2.

4 Time trend page: Selecting Years and Time trend

Time trend analysis is the first step before the user continues with regression or scenario analysis. Time trend analysis is not offered as an independent analysis. This explains that the user cannot retrieve analyst settings (see last remark previous chapter) and the user cannot view any results without entering the regression or scenario analysis.

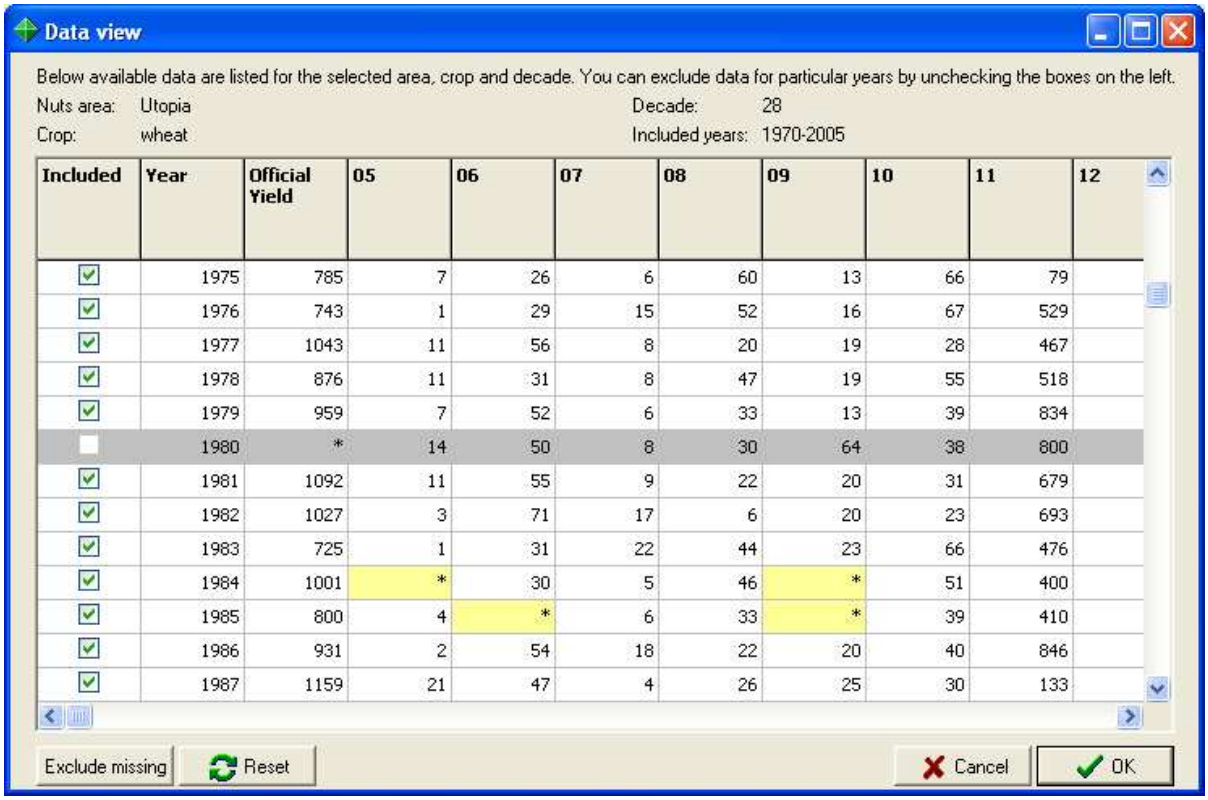
4.1 Selection of years

After the Area, Crop and Period are selected, the time trend page will be active. The time trend page initially displays the Start and End year for which there are yield data in the database. The default Target year, for which a prediction will be made, is the End year plus one. These values can be modified by the user. The only restriction is that there must be at least four years from Start to End year.



The user is informed about the time span covered by the years in the database, the number of years for which crop yield data are available in the database, and the number of excluded years. Years can be excluded by setting the Start and/or End

year or by opening the View Data Window. The following screen is opened by pressing the "View Data / Exclude more years" button.



Included	Year	Official Yield	05	06	07	08	09	10	11	12
☑	1975	785	7	26	6	60	13	66	79	
☑	1976	743	1	29	15	52	16	67	529	
☑	1977	1043	11	56	8	20	19	28	467	
☑	1978	876	11	31	8	47	19	55	518	
☑	1979	959	7	52	6	33	13	39	834	
☐	1980	*	14	50	8	30	64	38	800	
☑	1981	1092	11	55	9	22	20	31	679	
☑	1982	1027	3	71	17	6	20	23	693	
☑	1983	725	1	31	22	44	23	66	476	
☑	1984	1001	*	30	5	46	*	51	400	
☑	1985	800	4	*	6	33	*	39	410	
☑	1986	931	2	54	18	22	20	40	846	
☑	1987	1159	21	47	4	26	25	30	133	

Years with missing Official Yields are always displayed in grey and missing data in the indicators are highlighted in yellow. Individual years can be excluded or included by checkboxes on the left; years which are excluded are highlighted in grey. The checkboxes for years with missing Official yields, and for years outside the time span defined by the Start and End year, are disabled and also highlighted in grey. The target year is indicated by the word "Target" in the checkbox column. All years with missing values in any of the columns can be excluded by employing the "Exclude missing" button. All years, except the ones with missing Official yields, can be included by using the "Reset" button. The OK button must be used to confirm the changes made, or you can press the Cancel button to leave the Data View without making any changes. The Time trend page will be updated if necessary.

Note that the Official yields depend on the year only and are not affected by the decade selector on the opening screen. The indicator values are however specific for year and decade. In the installed sample database indicators are available for all decades of the growing season for a selected crop. In the case of the CGMS indicators the end values at maturity are stored in the data base for the remaining

decades of the year after the growing season. The database does not contain CGMS indicator values for decades preceding the decade of sowing.

4.2 Selection of the appropriate time trend

The bottom part of the Time trend page can be used to select the time trend which will be included in every regression model and every scenario model. To help the user to select an appropriate time trend, a graph of yield versus year is displayed along with a linear time trend in green and a quadratic time trend in blue. The displayed years are from Start to End year and manually excluded years, for which yields are available, are displayed by red crosses. Only the included years (black crosses in the year-yield diagram) are used for fitting the shown linear and quadratic time trend.

Five choices are available to specify or automatically select the time trend to be used in every regression model:

1. None – no time trend will be included.
2. Linear – a linear time trend will be included in every model, regardless of its significance.
3. Quadratic – a quadratic time trend will be included in every model, regardless of its significance.
4. Automatic testing up to linear – a linear time trend will be included only when it is significant. When it is not significant no time trend will be included.
5. Automatic testing up to quadratic – the time trend is determined by means of backward elimination (MPV 223). When the quadratic term, corrected for a linear term, is significant the resulting time trend is quadratic. In case the quadratic term is not significant but the linear term is, the resulting time trend is linear; when both are not significant no time trend will be included.

The selected time trend is always displayed on the right bottom of the time trend page.

The automatic testing methods employ a significance level which can be modified by the user. The spin buttons next to the significance level can be used to specify the nearest preset value, but you can also manually enter a value. For example, it is possible that at a significance level of 0.025 no trend is found, while for a significance level of 0.050 a linear trend is found. Note that automatic testing of the time trend is done without taking any of the indicators into account. Of course, the time trend is determined on the basis of the years included in the analysis; i.e. in case any years were excluded, it is irrelevant here what the reason was for doing so.

Finally the user can employ a logarithmic transform of the year. This can be used as an alternative for a linear or quadratic model. The logarithmic transform uses an offset which can be set by the user. So instead of using Log(year) as explanatory variable, Log(year – offset) is used. The offset is shown in the input box next to the

one showing the transformation for year. The maximum value for the offset equals the starting year minus 5. For example, if 1976 has been selected as the start year, the offset can not be larger than 1971.

4.3 Testing the trend

On the right of the time trend graph a list of p values for testing the linear and quadratic time trend is displayed. The p values for the quadratic trend are again corrected for a linear time trend. The top p values are for the time period from Start to End year, as selected by the user or selected by default. Underneath p values are given for time windows of decreasing length with the End year fixed and the Start year becoming one year later in every row further down. This enables the user to quickly select a different period, e.g. a period with a very significant linear time trend. Selection of a different period can be done by employing the radio buttons to the left of the p values. This will update the value of the Start year, the number of excluded years, the graphical display and the selected time trend displayed on the right bottom of the page. Note that clicking a radio button overrides the chosen model for time trend in the list down box. The choice made there can be re activated by clicking the Re apply button next to the Significance Level. This will also update the list of p values.

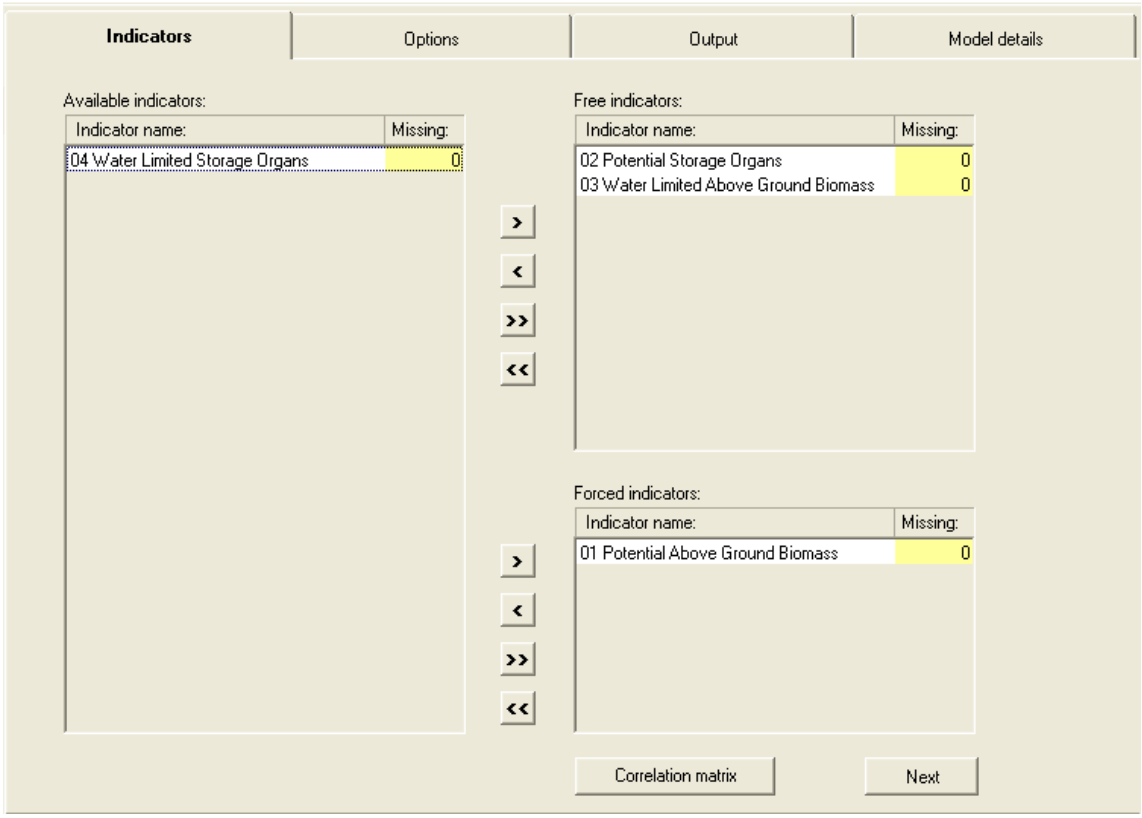
4.4 Analyzing the trend model

The application of the full regression model or scenario model requires the selection of a trend and one or more indicators. However, the user can analyze the trend only, without taking into account the indicators in the model. In this case the user should first select the regression or scenario analysis page and next navigate to the output page, by pressing each of the Next buttons on the bottom right of each of the successive pages: Time trend, Indicators, Options. Once the trend model has been selected in the output page the model details page will be updated.

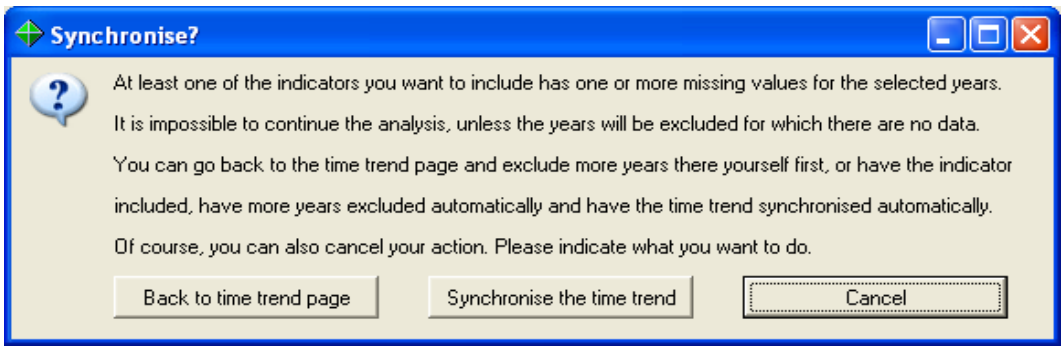
5 Indicators page: Selecting Indicators to Include

When the time period and time trend are selected in the Time trend page, the user can move on to Regression analysis or Scenario analysis. In any case, the user will have to continue on the Indicators page. This page presents the user with the available indicators.

For the mean time, we assume that the user will move on to Regression analysis. Available indicators can be moved to and from the Free or Forced list by selecting one or more indicators and then pressing the arrow buttons. Forced indicators will be included in every regression model, like the linear or quadratic time trend, while Free indicators are either excluded or included. In the example below indicator 01 is selected as Forced, while indicators 02 and 03 are chosen as Free. Indicator 04 will not be used in any of the fitted regression models.



Care must be taken when there are missing values in the indicators. The number of missing values for each indicator is displayed along with the indicator name. When an indicator with one or more missing values is selected as Free or Forced, the years with these missing values have to be excluded from the regression analysis. To make the user aware of this, a dialog appears:



The user is therefore presented with three choices:

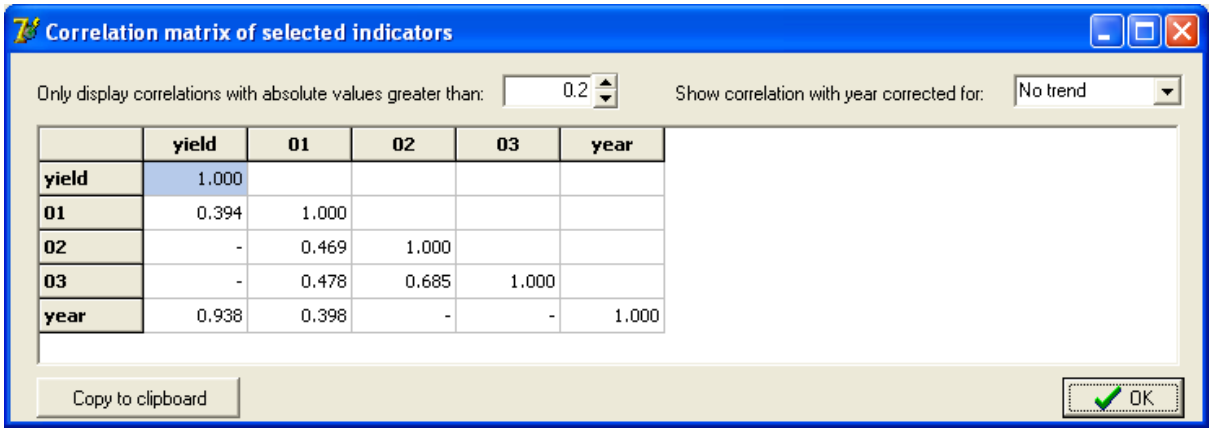
1. To go back to the Time trend page and the Data view and to exclude years there manually;
2. To have the years for which there are missing data excluded and to synchronise the timetrend automatically;
3. To cancel the current action of including of the indicators and to reconsider.

If the user chooses to synchronize the time trend automatically, an informational message appears in the log window with a list of the excluded years. Of course the indicators selected earlier are moved to the right side of the page and the column with missing years set to zero. In addition, the number of missing values for the other indicators still remaining on the left side is also updated. When automatic testing for time trend is requested on the time trend page, the excluding of extra years may also imply that a different time trend is selected. As mentioned, an informational message is shown with a list of excluded years, when this option is selected. It is a known bug that if there are years for which there are no yield data available, those years may or may not be included in those lists.

When there are many missing values, the user might want to go back to the Data View window to take a careful look at the data.

A correlation matrix can be viewed for the selected Free and Forced indicators by pressing the "Correlation matrix" button. Only absolute correlations above an adjustable threshold are displayed to aid the search for large correlations. The indicators are represented by their numbers for concise display of the correlation matrix. However the indicator names appear as tooltips when the mouse is moved over the indicator number. By default correlations with year are also given. The user can also request correlations corrected for a linear or a quadratic time trend. Such correlations are called partial correlations, and they are useful when a time trend is included in every model. For instance when a linear time trend is chosen, the

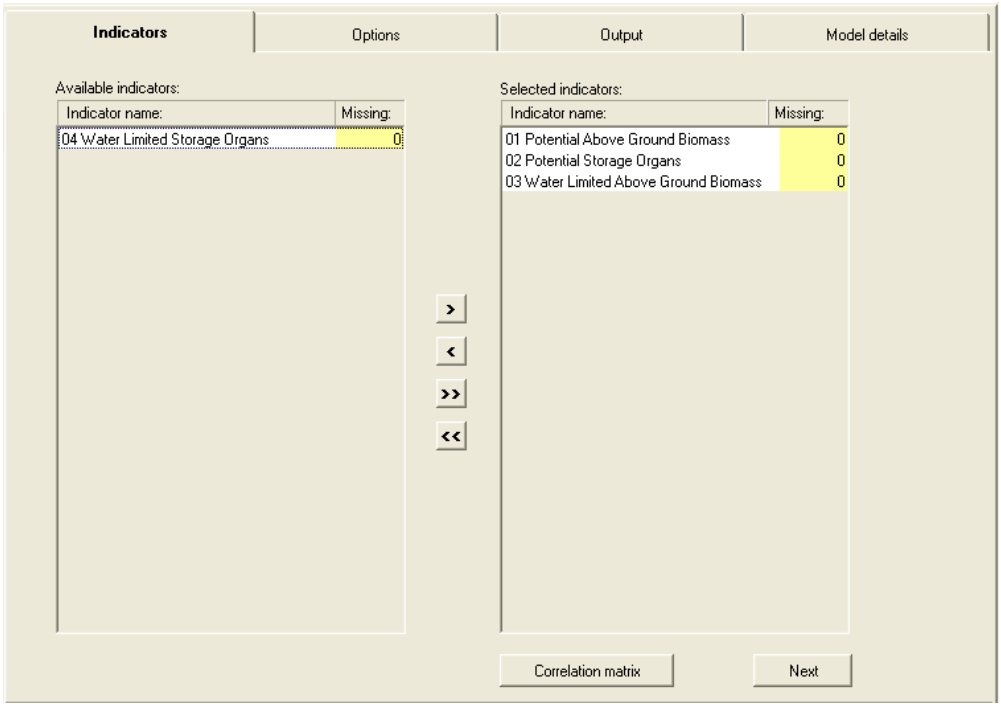
indicator with the highest partial correlation (i.e. corrected for a linear trend) with yield is the most important single indicator (MPV 310). Note that when partial correlations are requested, correlations with year are not displayed because they are not defined.



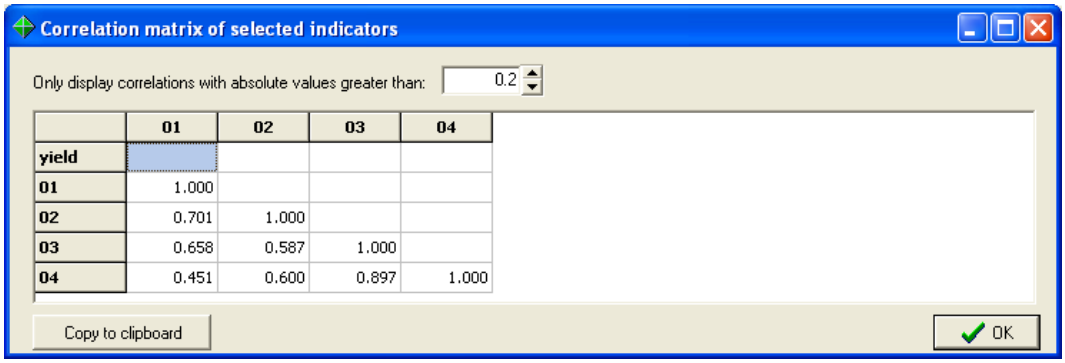
	yield	01	02	03	year
yield	1.000				
01	0.394	1.000			
02	-	0.469	1.000		
03	-	0.478	0.685	1.000	
year	0.938	0.398	-	-	1.000

The correlation matrix can be copied to the clipboard for inclusion in a document. The full names of the indicators can be retrieved from the file indicatornames.txt which is located in C:\Documents and Settings\All Users Application Data\Alterra\CgmsStatTool.

In the case of Scenario analysis, the user will be given similar options. However, indicators are not included into the model as such, but can be included for use in the principal component analysis. Moreover, no difference is made as to whether variables are free or forced. An example is shown below.



In the case of Scenario analysis, the window showing the correlation matrix looks like this:

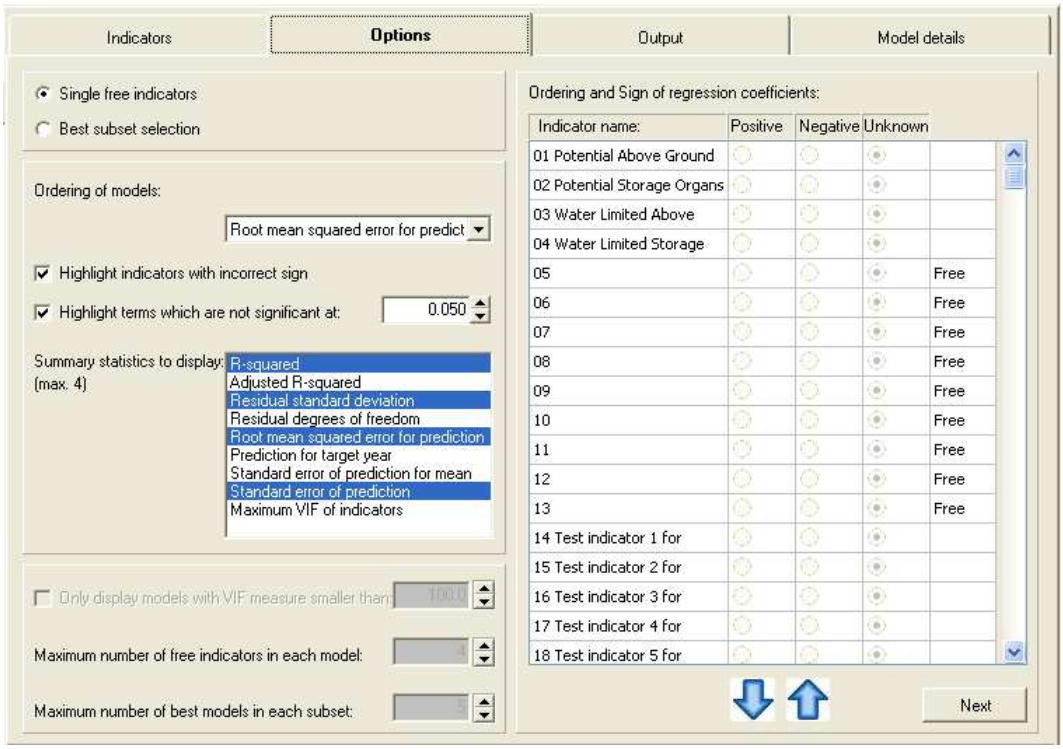


	01	02	03	04
yield				
01	1.000			
02	0.701	1.000		
03	0.658	0.587	1.000	
04	0.451	0.600	0.897	1.000

Note that the correlation matrix activated through the regression analysis exclude the target year while the correlation matrix of the scenario analysis does include the target year.

6 Options page: Setting Options for Output

After indicators are selected the user can move on to the Options page. First we assume that the user will move to the options of the regression analysis. The main choice here is between the "Single free indicators" method and the method of "Best subset selection". The single free indicator method only fits models with one Free indicator and the model without any Free indicators. Every model will encompass the Forced indicators and the chosen time trend. So when there are 5 free indicators, only results for 6 models will be presented, the 6th model being the model without a free indicator. The method of "Best subset selection" will display the best models with multiple indicators. Various options can be set to display specific summary statistics and to enhance the visual presentation of the fitted models. Some options are specific for the chosen method.



The fitted regression models will be displayed in the Output page, where every row represents a single model. The models are selected and ordered according to a single summary statistic which can be chosen by the user. The user can choose from R squared, Adjusted R squared, Root mean squared error of prediction, Standard error of prediction for mean or Standard error of prediction and Residual standard deviation. Mallows Cp statistic, which requires the fitting of a full model, is also available for best subset selection. Pros and cons for each statistic are given in the section Some Statistical Issues. A maximum of 4 summary statistics can be displayed

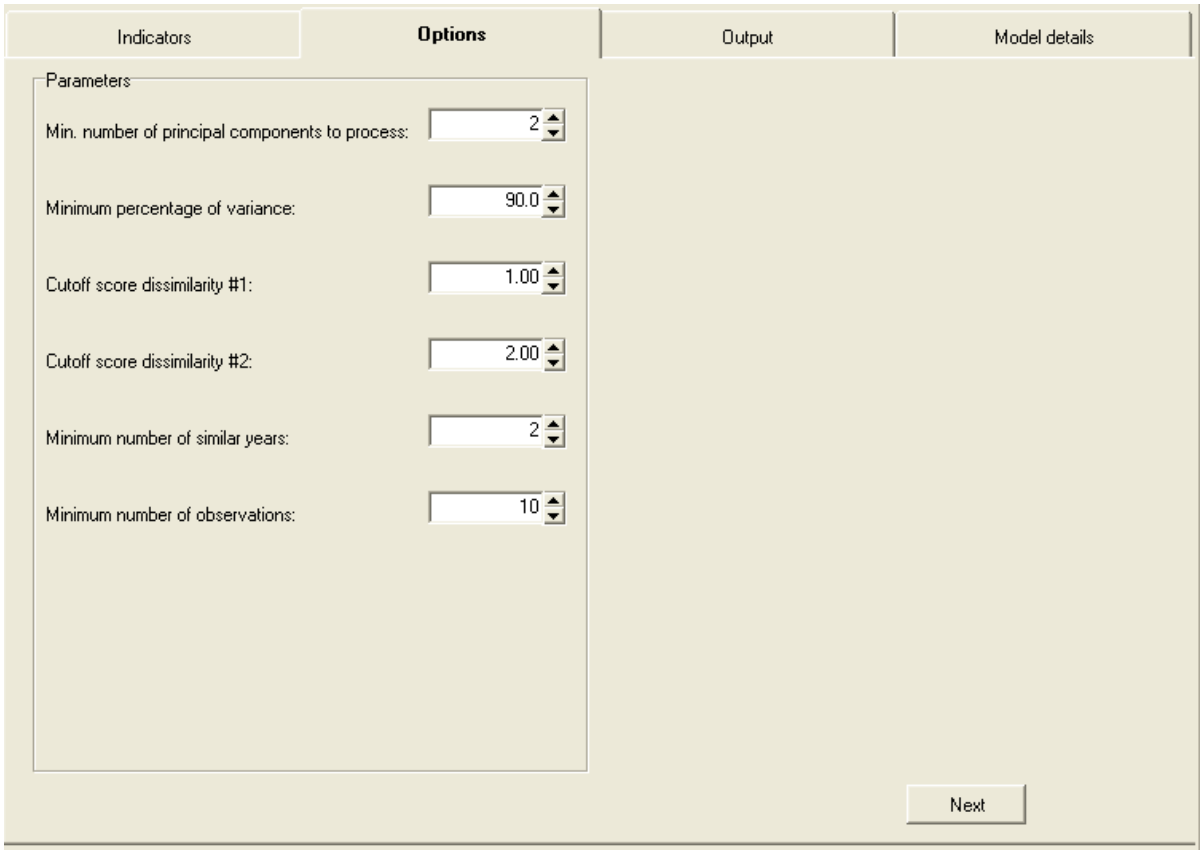
for every model. These include the already mentioned statistics, and also the Residual degrees of freedom, Prediction for target year and the Maximum of the Variance Inflation Factors (VIF) of the indicators. The use of the VIF is described in the Some Statistical Issues section.

The right part of the screen displays the list of indicators and whether they are chosen as Forced or Free. On the basis of experience, the user may expect a certain sign for the coefficients for at least some of the indicators – please read more on this in the section Some Statistical Issues. In such cases the user can set the expected sign of a regression coefficient by means of the radio buttons. This can be done for each indicator, irrespective of the question whether that indicator is selected or not. The sign can be Positive, Negative or Unknown. The column headers can be clicked to set all signs simultaneously. When the option "Highlight indicators with incorrect sign" is checked, all regression coefficients with the incorrect sign will be highlighted in the Output page. When the sign is set to Unknown no highlighting will occur. It is also possible to highlight indicators for which the corresponding estimate is not significant at an adjustable level. These options can help to select an appropriate regression model.

For the method of best subset selection, three extra options are available. The search for the best models can be subject to an optional constraint, set by the VIF option, on the degree of correlation permitted among the indicator variables. A higher value for the VIF measure allows models with a higher degree of correlation among the indicators; see the section 9: Some Statistical Issues. The maximum number of free indicators in every subset is limited by 4. This is to prevent the user to choose a model with many indicators. Note however that - although not recommended - models with many indicators can be fitted by selecting a lot of indicators as Forced. Finally the user can set the number of models to display for each subset.

The best subset algorithm requires that the full model, i.e. the model with time trend and all Forced and Free indicators, can be fitted. The full model can not be fitted when the number of included years is less than the number of regression coefficients to be estimated, or when indicators themselves are linearly related. This is called aliasing of Indicators. In that case, after fitting the time trend, the selected Forced and Free indicators are added subsequently to the model. Indicators which can not be fitted, either due to lack of sufficient years or due to linear relations among the indicators, are dropped from the list. The order in which indicators are added can be specified by the user by using the up and down arrows below the list of indicators. Indicators on top of the list are added first.

In the case of Scenario analysis, the Options page contains various options for making sure that the desired number of similar years is obtained, i.e. years which are similar to the selected target year.


Indicators
Options
Output
Model details
Parameters
Min. number of principal components to process:
2
Minimum percentage of variance:
90.0
Cutoff score dissimilarity #1:
1.00
Cutoff score dissimilarity #2:
2.00
Minimum number of similar years:
2
Minimum number of observations:
10
Next

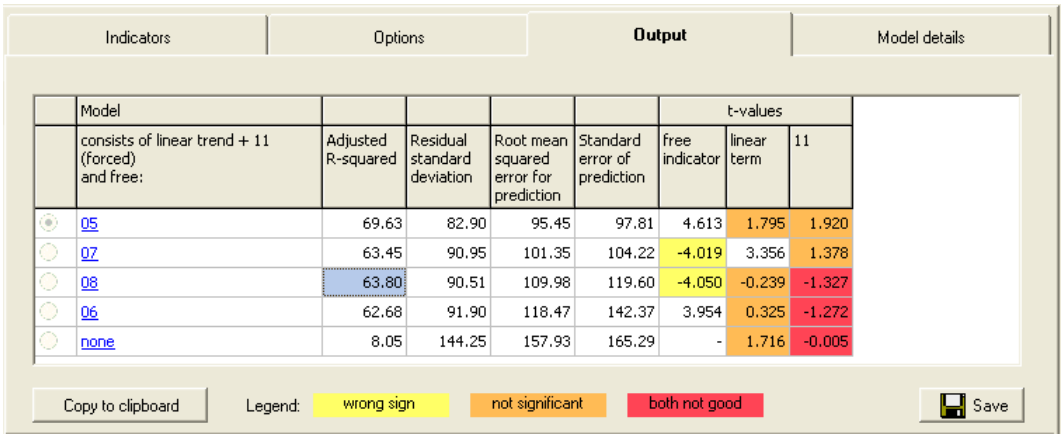
7 Output page: Viewing the results

Once all options are set in the Options page, the user can move on to the Output page.

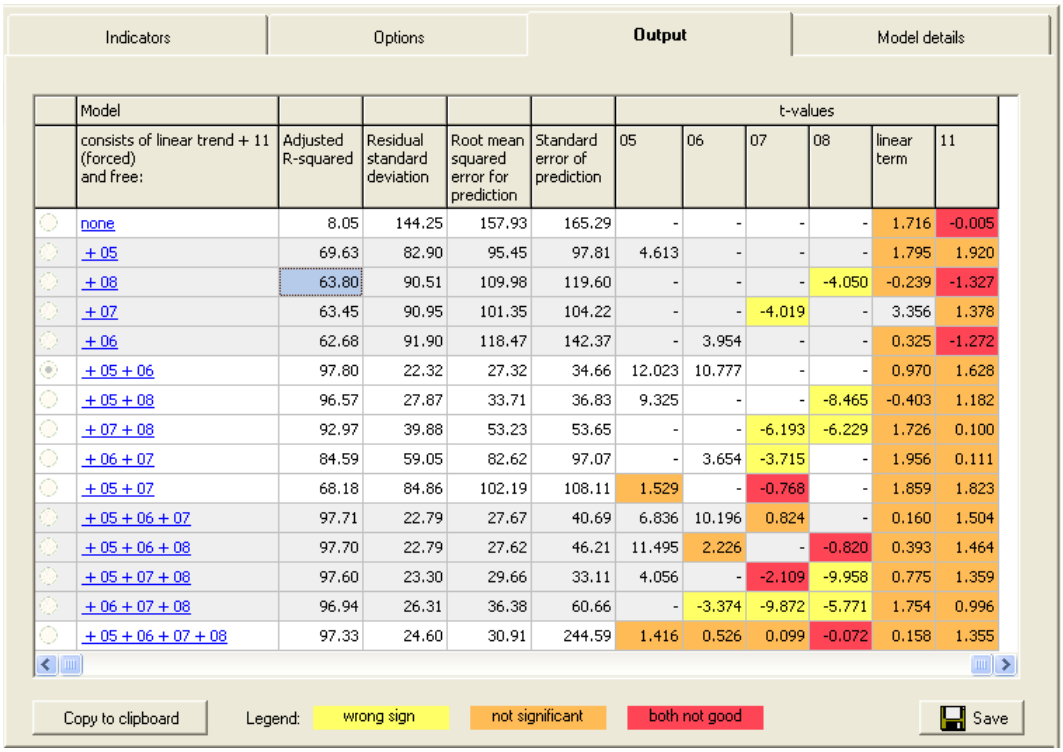
7.1 Regression Models

In the case of Regression analysis, the requested method of single indicators or best subset selection is applied and results of the fitted regression models are displayed in the Output page. As a hypothetical example, consider the Utopia dataset from 1975 to 1990, excluding years 1980, 1984 and 1985 for which there are missing values. Target year is 1991 and the first dekad of October is selected. A linear time trend was requested, indicator 11 was selected as Forced and indicators 05, 06, 07 and 08 were selected as Free. Furthermore both types of highlighting were requested, with a significance level of 5%, and the models must be ordered according to Root mean squared error of prediction. The sign of each regression coefficient was expected to be positive.

The output for the method of Single free indicators is given below. Every row represents a model and the second column lists which free indicator is included. The model indicated by "none" is without a free indicator. The header of the second column reminds the user of the chosen time trend and forced indicators. Next to the model are four summary statistics, indeed sorted according to the Root mean squared error for prediction. The column denoted by "free indicator" contains the t value for the regression coefficient of the associated free indicator. The "linear term" column lists the t value for the linear time trend, and the "11" column the t value for the forced indicator. The coloring of the t value indicates a wrong sign of the coefficient (yellow), a coefficient which is not significant (orange) or both not good (red). Clearly, indicator 11 is never significant and the user might want to drop 11 from the model. Although the adjusted R squared values of the top 4 models are quite similar, the models give very different predictions for the target year. Also note that the standard errors of prediction for the target year are quite large. The radio button in front of the first model indicates that this is the best model according to the chosen criterion. You can select a different model by clicking on the associated radio button.



The method of best subset selection presents the user with the following list of models.



The output is very similar to the output discussed before. The main difference is that more than one free indicator can enter a regression model. Every free indicator now has its own t value column with a value only when the associated free indicator is included in the model. The models are sorted according to the number of included free indicators, and within that by the requested summary statistic. The alternating color of the rows, white or light grey, is used to distinguish models with 1, 2, 3 or 4 free indicators. Clearly the models with 2 free indicators are better than the ones with only 1 indicator; and there is not much point in using a model with 3 or 4 indicators. The best model is the one with free indicators 05 and 06 with very significant t values; moreover the free regression coefficients have the correct sign. Note that for

almost all models neither the linear time trend nor the forced indicator 11 is significant. An analysis without a linear time trend and with indicator 11 removed might therefore be useful.

In case the user wants to reproduce the results shown on the Output page elsewhere, he or she can copy the content of the window to the clipboard, by clicking first on the button "Copy to clipboard", then paste into Excel or Word file. The full names of the indicators can be retrieved from the file indicatornames.txt which is located in in C:\Documents and Settings\All Users \Application Data\Alterra\CgmsStatTool.

The Save button at the bottom of the page can be used to save the results for the model indicated by the radio button. The results are saved to the database tables MODEL_REGR_INDICATIFS, MODEL_REGR_EXCL_YEARS and MODEL_REGR_INCL_INDICATORS. More details on the structure and content of these tables can be obtained from the section 10.2 and from Annex 1. The save button also saves all settings of the user interface (see section 10.2).

7.2 Scenario models

The output of a typical scenario analysis is given below. The example pertains to wheat in Portugal for the years 1975 to 2004 and with the year 2005 as the target year. Four CGMS indicators were included: Potential and Water Limited Above Ground Biomass and Potential and Water Limited Storage Organs. As can be seen, a quadratic time trend was selected. The top left section of the output page shows three statistics about principal components, explained variance and dissimilarity.

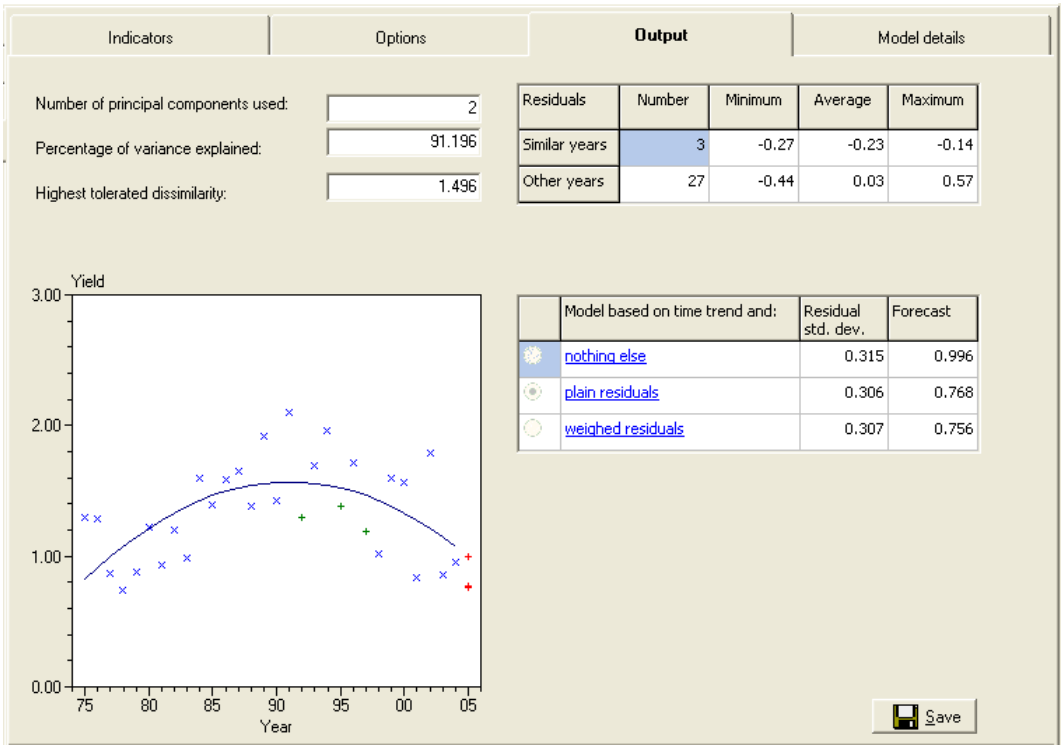
The scenario analysis involves a principal component analysis first on the indicator vectors. As many components are added as is necessary to at least explain a given minimum level of variance. The components are then used to calculate a distance matrix; to be exact: a matrix which contains squared Euclidian distances. Such distances indicate how far in the Euclidian space a year is located away from other years. A set of years is then selected - on the basis of this matrix - which are closely located to the target year: the similar years.

Subsequently, some calculations are done on the residuals of the similar years and the other years. On the right, the Output page shows the number of similar years and the number of other years – or non-similar years. In the same table, the minimum, average and maximum residuals are given for similar years and for the non-similar years respectively.

Furthermore, the Output page shows results of three predictions, based on different methods. The first prediction is based on the time trend alone, without making use of residuals – therefore marked as "based on nothing else". The remaining two predictions are indeed made by means of the residuals. The forecast according to the time trend is taken into account, but it is corrected with a quantity:

- an average of the residuals of the similar years – marked as "based on plain residuals"

- a weighed average of the residuals of the similar years – marked as "based on weighed residuals".



The mentioned weights are set to the inverse of the squared distances between each of the similar years and the target year. Weights are scaled by dividing each of them by their sum – i.e. they sum up to 1. In order to make sure that the residuals of years which are very similar to the target year would exert a too great influence on the prediction, a maximum was introduced for the unscaled weights – default value for this is 50.

From the three predictions, the one with the lowest residual standard deviation is selected. The residual standard deviation is the only criterion used for automatic selection from the three different models, although this is probably not the best criterion.

For the predictions which are derived from the mentioned residuals, the residual standard deviation is calculated by adding the sum of squares of the non-similar years to the sum of squares of the similar years. The latter is calculated by making predictions for each of the similar years in the same way as is done for the target year, i.e. a correction is calculated for each similar year on the basis of the other similar years. This is done for all the similar years, without calculating the distance matrix again. This means that the principal component analysis is done only once! The differences between the obtained predictions and the observed yields are then used to calculate the sum of squares of the similar years.

8 Model details page: Viewing Results of a Single Model

Details of a model can be obtained by clicking on the blue link for that model in the Output page. The details are then displayed in the Model details page which is opened automatically. One can Copy the details, or parts of the details, and then Paste them into Microsoft Word. The Model details page for a regression model includes the following five sections:

- Description of model
- Summary statistics
- Regression coefficients
- Case statistics
- Plots.

The model details page for scenario analysis, includes the following sections:

- Description of model
- Principal Component Analysis – parameters
- Principal Component Analysis – loadings
- Principal Component Analysis – statistics
- Cluster Analysis
- Analysis of residuals
- Summary statistics
- Prediction
- Case statistics

In the following, the above sections are dealt with in more details. For a higher level treatise of the statistical issues involved, see the next chapter entitled "Some Statistical Issues".

8.1 Description of model

This section lists the NUTS area, crop, decade, number of included years, start, end and target year. Also listed is whether the years have been transformed, the offset for the year effect and the time trend. For regression analysis an additional row is added which shows which indicators are included. For scenario analysis, this row indicates which of the models was selected: the one based on the time trend and "nothing else", also on the plain residuals or also on the "weighed residuals". Note that the years which are included in the analysis are listed in the Case statistics section.

8.2 Summary Statistics

For regression analysis, the following summary statistics, and the page in MPV where they are defined, are listed:

- R squared: the percentage variance explained (MPV 39)
- Adjusted R squared: the adjusted percentage variance explained (MPV 90)
- Residual Standard deviation: the square root of the residual mean square (MPV 23)
- Root mean squared error for prediction: define e(i) as the difference between the i th response and the predicted value for the i th response based on a model fit to the remaining observations, i.e. without the i th observations. This is sometimes called the PRESS residual or the leave one out residual. The Root mean squared error of prediction is the root of the mean value of all the squared e(i). This is similar to the PRESS statistic (MPV 153).
- Mallows Cp: a comparison of the residual mean square of the model under consideration, and the residual mean square of a full model, i.e. a model which includes all indicators as well as a time trend. Models, for which the Cp statistic is similar to the number of regression coefficients in the model, are considered to be good models (MPV 299).
- Maximum of VIF: the variance inflation factor of a regression coefficient is a measure of the degree of correlation between the associated indicator and the remaining indicators. Large correlations are equivalent to large variance inflation factor (MPV 337). This summary statistic is the maximum of the inflation factors for the indicators, i.e. the inflation factors for the constant and the time trend effects are not taken into account.
- Prediction for target year: the mean yield prediction for the target year according to the regression model (MPV 34).
- Standard error of prediction (Mean): the standard error of the mean yield prediction for the target year (MPV 34).
- Standard error of prediction (New): the standard error of the individual yield prediction for the target year (MPV 37). The following equality holds: SeNew2 = SeMean2 + ResidualStandardDeviation2.
- Residual degrees of freedom: the number of degrees of freedom for residual (MPV 23)

For scenario analysis, only the following statistics are listed:

- R squared: the percentage variance explained (MPV 39)
- Residual standard deviation: the square root of the residual mean square
- Prediction for target year: the yield prediction for the target year according to the time trend and corrected with a linear combination of residuals pertaining to the similar years
- Prediction for target year, pessimistic value: the yield prediction for the target year according to the time trend corrected with the lowest residual available for the similar years

- Prediction for target year, optimistic value: the yield prediction for the target year according to the time trend corrected with the highest residual available for the similar years
- Residual degrees of freedom: the number of degrees of freedom for residual (MPV 23)

8.3 Coefficients

For regression analysis, this section lists the estimates of the regression coefficients, their estimated standard errors, the associated t values and two sided p values. To increase numerical precision, the regression coefficient for the linear time trend is for (year offset) rather than year itself. The offset was fixed at 1970. Likewise, the regression coefficient for the quadratic time trend is for $(\text{year offset})^2$. In case of a logarithmic transformation of the years, 1970 is also the default value for the offset, but in such cases the offset can be changed by the user to a certain extent.
The p values are calculated by means of a Student distribution with degrees of freedom equal to the residual degrees of freedom. Finally the individual variance inflation factors (VIF) are listed. The VIF of an indicator is a measure of the degree of correlation between that indicator and the remaining indicators and time trend, if any, in the model.

For scenario analysis, this section only lists the estimates of the time trend coefficients.

8.4 Case statistics

For each included year the following case statistics are listed:

- The yield.
- The fitted value according to the regression model. Note that the fitted value in the last row equals the prediction for the target year.
- The ordinary residual which is the difference between the yield and the fitted value.

For regression analysis, the following case statistics are shown as well:

- The leverage (MPV 209). Observations with large leverages are potentially influential because these observations have remote indicator values as compared to the rest of the observations. The mean of all leverages equals p/n in which p is the number of regression coefficients and n the number of observations. Leverages larger than 2p/n are traditionally considered as high leverage points. The leverage of the target year is of special interest because a target year with remote indicators will have a large standard error of prediction.
- The influence on the target year prediction. This is the difference between the predicted value for the target year based on the full dataset and the

predicted value when the i the observation is removed from the dataset. This is similar to the DFFITS statistics (MPV 214).

8.5 Results specific for scenario analysis

In the case of scenario analysis, there are three sections which show the results of the Principal Component Analysis (PCA). In the first place, the following parameters are of importance:

- Number of principal components
- Percentage of explained variance
- Number of observations used.

Secondly, one may like to know in which way each of the components is composed of the original indicators. The components are always linear combinations of those original indicators and the so-called loadings are therefore the relevant coefficients.

Thirdly, the components are always standardized with mean and standard deviation, so that they have an expectation of 0 (= zero) and a variance of 1 (= one). The resulting factors can then be compared more easily and scores plotted against similar axes. One may like to know which mean and standard deviation were used in the standardization, so these are shown in this section.

After the Principal Component Analysis, Cluster Analysis is done. Relevant in this context are:

- the cutoff scores used
- maximum Euclidian distance to the target year
- the number of similar years found

In the section Analysis of Residuals, similar years are placed versus the non-similar years by showing some simple statistics of the two sets of years.

Finally, the section indicated as "Prediction" shows the following per similar year:

- the Euclidian distance to the target year
- the weight for the similar year
- the residual for the similar year
- the contribution of the similar year to the correction; i.e. the product of weight and residual.

8.6 Plots for diagnosing the model

Two plots of case statistics can be used to check various aspects of the model:

- Observed (o) and fitted (x) values versus year. This plot is similar to the plot on the Time trend page, except that for excluded years no values are displayed at all. The plot can be used to check the observed and fitted yields.

Note that for the target year only the predicted value is displayed; the prediction can thus be compared with the observed yields.

- Residuals versus fitted values, with an added horizontal line at y=0. The residuals should be evenly distributed for every fitted value. For example, when the residuals spread more for higher fitted values this indicates that the variance of the observations is not homogeneous (MPV 140)

For regression analysis, the following plots are shown:

- Normal probability plot of the studentized residuals (MPV 134) with a straight line to aid interpretation of the plot. The expected normal scores Φ^{-1} [(i 0.375) / (n+0.25)] are used, see McCullagh and Nelder (1989), and the points must roughly lie on a straight line. Although normality in itself is not an important assumption, a normal probability plot is still useful to identify outliers (MPV 138).
- Ordinary residuals versus year, with added horizontal lines at y=0 and at two sided cut off values with a p value of 0.95. The residuals should not have a relationship with year. A linear or quadratic plot might indicate that a time trend should be added to the model. The plot can also reveal that the variance is changing with time, or that residuals are correlated (MPV 146).
- Leverage versus year, with an added horizontal line at 2p/n. This plot can be used to identify observations that are potentially influential.
- Influence on the target prediction versus year, with horizontal lines at 2 x √ (p/n) x √ (LevN x RMS), where LevN is the leverage for the target year and RMS is the residual mean square. This plot shows the effect of deleting individual observations on the predicted value for the target year. A point can be influential on the regression model as a whole, but have a small target prediction residual. See MPV 214 for the cut off value.

For scenario analysis, the following plots are shown:

- Dendrogram: a tree diagram used to illustrate the arrangement of clusters; to obtain the clusters, the clustering algorithm developed by Ward was used (Ward, 1963) with Euclidean distance as the type of distance measurement.
- One or more plots of factor scores: in case of 2 principal components, only one plot is shown, in case of more components three plots are shown. Ovals are drawn in the plots, indicating the two cutoff distances used in the analysis. Those ovals are really circles with radii equal to the square roots of the respective cutoff values, i.e. the factors all have expectation 0 and variance 1.
- A screeplot – sometimes also called eigenvalues diagram – is a graph used in the context of factor analysis showing the eigenvalues of the candidates in descending order of magnitude.

9 Some Statistical Issues

In this chapter, the statistical background is explained of the model selection as is facilitated by the CgmsStatTool. In section 9.1, various summary statistics for use in regression analysis are described, with their advantages and disadvantages for arriving at a stable prediction model. In section 9.2, it is described what the "best subset" algorithm involves and why it was chosen for the tool. Section 9.3, 9.4 and 9.5 describe some of the pitfalls related to model selection, particularly for regression analysis.

Besides statistical principles, past experience is also relevant for model selection. Past experience has shown that, within the CGMS framework, time trend is generally the most important indicator. It is therefore that the user must specify a time trend (none, linear or quadratic) before any other element is added to the model. In case of regression analysis, the expected sign of the regression coefficients for some indicators may be known, also due to past experience. Estimated coefficients with a wrong sign can therefore be highlighted as an indication that the model might not be correct.

In general, one should always try regression analysis first in order to arrive at a stable prediction model. Prediction models based on regression analysis are particularly useful for large areas, for years with more or less usual weather. Adverse weather conditions rarely occur over larger areas and do not usually also affect larger areas in the same unfavourable way. A dry period may e.g. cause yield reductions on certain types of soil, whereas it may be favourable on other soils which are generally subject to high groundwater tables. Therefore drastic yield reductions and even complete crop failures are often averaged out easily.

For countries with a rather even climate, the time trend alone already gives a good fit and adding an extra indicator to the regression model often does not improve the model much. On the other hand, for countries with a rather whimsical climate adding an extra indicator may improve the model considerably.

Scenario analysis is expected to be particularly useful for smaller areas, for years with exceptional weather. Section 8.6 contains more considerations on scenario analysis.

9.1 Selection of the Best Model

There are various methods for choosing a regression model when there are many indicators. Commonly used methods are forward selection, backward elimination and stepwise regression (MPV 310). However these methods result in only one model and alternative models, with an equivalent or even better fit, are easily overlooked. Moreover, the particular indicators that affect the response and the directions of their effects are of intrinsic interest and then selection of just one well fitting model

is unsatisfactory and possibly misleading. Therefore both the single indicators and the best subset selection methods present the user with several models. However, any selection method should be used with caution, especially when the number of indicators is large in comparison with the number of observations. In this case uncritical model selection might lead to models which appear to have a lot of explanatory power, but contain noise variables only, see e.g. Flack and Chang (1987). Indicators should therefore not be selected on the basis of a statistical analysis alone. Experience with the intended use of the model, i.e. prediction for a target year, is therefore very important.

As mentioned above, the sign of coefficients may be known from past experience. Wrong signs can be due to collinearity among the indicators, or because other important indicators have not been included in the model (MPV 120). Because the principal aim of the CGMS statistical system is to provide a reliable prediction for the target year, it should be noted that models with as few indicators as possible generally have better predictive power than models with many indicators. This can easily be understood when one realises that the more indicators are included in the model, the more coefficients will have to be estimated and therefore the more uncertainty is built into the model. The significance level of a regression coefficient indicates whether the corresponding indicator is necessary or not. Non significant indicators can therefore be highlighted. So using highlighting both for sign and significance can be employed to select a model for which none of the regression coefficients is highlighted, although this might be too restrictive in practice.

Several summary statistics can be used to initially select a best model (MPV 296). Any statistic will generally be overoptimistic because a lot of models are fitted in the process, especially for best subset selection, and the best model can be a stroke of luck. In any case, R squared is not a good criterion to select the best model because it will always select a model with the maximum allowed number of indicators. That is because R squared always improves by adding an indicator to a model. To put this another way, there is no penalty for adding an indicator. When Adjusted R squared or Mallows Cp is used there is a penalty for adding an indicator. Adjusted R squared improves when the absolute t value of the added indicator is larger than 1, while Cp improves when the absolute t value is larger than $\sqrt{2}$. Clearly Cp is the more conservative criterion and will tend to select models with fewer indicators as compared to R squared and R squared adjusted. The Cp statistic also has another rationale. Assume that the full model, i.e. the model with all indicators, provides a good estimate of the residual variance. Then the expected value of the Cp statistic for a smaller model with no bias equals the number of regression coefficients. So Cp can be compared to the number of coefficients (including the constant). However when the number of indicators is large as compared to the number of observations, the full model will often overestimate the residual variation and consequently the values of the Cp statistic will be small (MPV 300). The Root mean squared error of prediction has an intuitive appeal for the CGMS Statistical Tool because it specifically aims at small prediction errors. It is commonly used for model selection. The standard error of prediction for the target year is an unconventional summary statistic. It is not known whether this criterion provides stable and reliable predictions. The standard

error of prediction for a new observation seems more appropriate than the standard error for the mean because the aim of CgmsStatTool is to predict a new observation.

The above remarks can be summarized as follows. Incorporating knowledge and experience in the process of selecting a best regression model is extremely important. When the number of observations is small as compared to the number of indicators, careful model selection is crucial. In such a case, automatic methods with all indicators might well select a model which appears to be very good but will have low predictive power. For that reason, a careful pre selection of indicators is necessary. The sign and significance of regression coefficients might provide clues as to which models are good.

9.2 The best subset model may not always be the best

The interface suggests that the method of best subset selection will always display the best models according to every summary statistic. However the algorithm used selects the best 40 models in every subset according to R squared. This is by definition equivalent to the best models according to R squared adjusted and Mallows Cp, but not always equivalent to the best models according to other summary statistics. For the 40 models thus selected, the other summary statistics are calculated and the models are sorted according to the requested statistic. Since a maximum of 10 models is displayed for every subset, it is likely, though not necessary, that these are the best models for every criterion. In theory however the best model according to one of the other statistics can be missed by the algorithm.

So why not use another algorithm? One could of course fit every possible subset in turn to select the best model according to any criterion. However with 20 free indicators there are 4845 possible models with maximally 4 indicators; with 30 free indicators there are even 31930 possible models. Clearly the computational burden would then be enormous. The algorithm used in CgmsStatTool however does not explicitly fit all models, but uses a branch and bound algorithm to find the best models. This implies that it is very efficient even for large numbers of indicators. It was therefore decided to use this algorithm. The minor disadvantage of possibly not selecting the best model according to any criterion is taken for granted.

The best subset algorithm used is the 1981 double precision version of a branch and bound algorithm for subset selection developed by Furnival and Wilson. This Fortran algorithm was obtained in 1982 by personel communication with Furnival and Wilson, see Ter Braak and Groeneveld (1982). It was claimed that the 1981 version is twice as fast as the 1974 version (Furnival and Wilson, 1974) and requires much less storage. From 1992 onwards, the best subset algorithm is also made available in Genstat by means of procedure RSELECT, see Goedhart (2005).

9.3 Multicollinearity and Variance Inflation Factors

Indicators are said to be multicollinear, or collinear, when there are near linear dependencies among the indicators (MPV 117 and 325). Near linear dependence can occur when there are many indicators as compared to the number of observations, or in case some indicators essentially measure the same aspect as other indicators. An simple example is the mean temperature between 7 and 19 hours and the 24 hours daily temperature.

Collinearity results in very large variances and covariances for the estimators of the regression coefficients. This implies that a small perturbation of the data might give very different estimated regression coefficients. Regression models with collinear indicators may therefore perform poorly when used for prediction purposes. One way to identify collinear indicators is that the sign of the estimated regression coefficients is wrong, or that the estimate is very large. This can simply be explained by the following example. Suppose that a response Y is, except for random error, linearly related to an indicator Z1 in the following way Y = 1 + 2xZ1. Suppose further that another indicator Z2 almost measures the same as indicator Z1, i.e. Z2 ≈ Z1. In that case Y = 1 + 2xZ2 is more or less equivalent to the model with Z1. But Y = 1 + 1xZ1 + 1xZ2 is also alike, and so is Y = 1 + 100xZ1 – 98xZ2. It turns out that all models for which the sum of the regression coefficients for Z1 and Z2 equals 2 are equivalent. As a result the estimated regression coefficients can be almost anything, depending on the specific observed values for Y, Z1 and Z2. Consequently, the variances of the estimates are generally large.

A useful measure of collinearity is the variance inflation factor (VIF). Define R_j^2 as the R squared value of a regression model of indicator j on all the remaining indicators. The VIF of indicator j is then defined as $100 / (100 \quad R_j^2)^{-1}$. If indicator j is unrelated to the remaining indicators, R_j^2 is small and the corresponding VIF will be close to unity. However when indicator j is linearly related to some subset of the remaining indicators, R_j^2 will be close to 100 and the VIF will be very large. It can be shown that the estimated variance of the regression coefficient for indicator j equals $(s^2 \, VIF_j)$, where s^2 equals the estimate of the residual variance. Clearly one or more large VIF values is indicative for collinearity. MPV claim that "practical experience indicates that if any of the VIFs exceeds 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity".

The maximum VIF of all indicators is thus a good summary statistic for a regression model. Note that in calculating the individual VIFs the time trend model is also taken into account. However since year and $year^2$ are themselves heavily correlated, even when an offset is first subtracted, the VIFs of the linear and quadratic time trend can be very large. This can be easily remedied by using orthogonal polynomials in which case the VIFs are equals to 1. This implies that VIFs for polynomial models are not very informative. Consequently, the maximum VIF of the regression model does not take the VIFs of the linear and quadratic time trend into account.

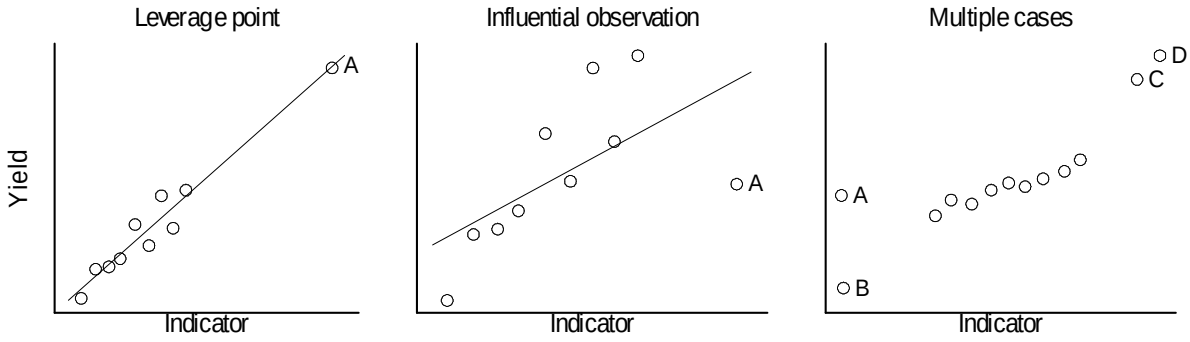
The method of best subset selection has an option to only select models with a maximum VIF smaller than an adjustable value. This can be useful when all the best models have very large maximum VIFs. By specifying a smaller maximum VIF value a deeper search for the best model can be enforced.

9.4 Regression Diagnostics and Case Statistics

Something is "wrong" with most regression models: a high leverage point, a large residual, a VIF which is somewhat large, a normal probability plot which is not a straight line, residuals that tend to become larger when the fitted value increases, an observation that has a large effect on the predicted value for the target year. Automatic removal of observations to remedy the problem is usually not a good idea. Deleting such points does improve the fit, but might result in a false sense of precision of the estimated regression model and thus of the prediction for the target year. Instead it is important to first find out why a certain problem occurs, and then take an appropriate action. It should be noted that the cut off values in the diagnostic graphs usually identify more observations than the data analyst is willing to verify. This is especially the case in small samples. Moreover, after removal of one point, sometimes another observation may suddenly stand out as problematic. In the discussion below a distinction is made between case statistics for outliers and for influential observations.

Outliers are observations with a response which is not typical of the rest of the data. They can generally be identified by looking at residuals, although multiple outliers might mask each other. It is important to try to understand why an outlier has occurred. Perhaps the yield of a crop is incorrectly recorded in a certain year, or the yield is very low because of extreme weather conditions which are not representative for the year for which a prediction has to be made. In such cases there is evidence that something is wrong with the observation and it can be safely removed from the dataset. However when there is only statistical evidence that the observation is outlying, one should think twice before deleting the observation. MPV (154) note that "there should be strong non-statistical evidence that the outlier is a bad value before it is discarded". Because the main aim of CgmsStatTool is to provide a prediction for the target year, the Influence on the target year prediction might be helpful in deciding whether to remove an observation or not. When there is only a small effect of removal of an outlier on the predicted value, one might want to keep the observation in the dataset.

Influential observations have a more than average effect on summary statistics and /or on the estimates of the regression coefficients. One can distinguish between leverage points and influential points as shown in the Figure below.



A leverage points has remote values for the indicators, but the regression line does not change much by deletion of the leverage point. Such a point can be identified by a large leverage and a small residual, and also a small target prediction residual. An influential observation on the other hand also has a large leverage, but now both the residual and the target prediction residual is also large. Discarding of influential observations is therefore always worth considering, while leverage points can be possibly left alone. Whether an influential observation should be discarded is analogous to the treatment of outliers. MPV (218) note that "if analysis reveals that an influential point is a valid observation, then there is no justification for its removal".

All regression diagnostics implemented in CgmsStatTool are calculated for individual observations. However Cook and Weisberg (1982) note that "it can happen that a group of observations will be influential, but this influence can go undetected when cases are examined individually". They illustrate this with the Figure on the right which is displayed above. If Point C or D is deleted, the fitted regression will change very little, while if both are deleted the estimates may be very different. Conversely, if A or B is deleted the fitted line will change, but if both are deleted the fitted line will stay about the same. This is simple to detect when there is only one indicator, but very hard when there are multiple indicators. Various suggestions have been made to identify groups of influential observations, among which certain forms of cluster analysis (MPV 217), but none of these have been implemented.

9.5 Perfect Fit and Aliasing of indicators

When a chosen time trend model whether constant, linear or quadratic provides a perfect fit to the yield data, the residual mean square of the regression model is zero. In this case a warning will be issued in the log window as soon as the data are processed for display on the Time trend page. Moreover the Time trend page will not display p values for the linear and quadratic timetrend. Also on the Output page the p values for linear and quadratic effects will be denoted by "alias". In that exceptional case no prediction for the target year is provided. This is also the case when a model with one or more forced indicators fits the data exactly. So only when a model with a time trend and forced predictors does not fit the data perfectly, a prediction for the target year is calculated.

When indicators are perfectly linearly related, they can not enter the same regression model. This can be either due to linear relations by definition or by chance, or by too few observations in combination with too many indicators. When the sample size is smaller than the number of indicators to be included in the regression model, the indicators are aliased by definition. Every effort has been made to ensure that the software handles aliasing in a correct and understandable way. The single free indicator method and the best subset selection method deal with aliased indicators in a different way. The single free indicator will display all indicators in the Output Tab, and aliased indicators can be identified by the word "alias" in the t value columns. A detailed analysis of such a model, by clicking on the blue link of the model, will have missing values for the corresponding regression coefficients. The best subset selection method on the other hand will first remove all aliased indicators from the indicator list, and proceed with the remaining indicators. This is because the algorithm requires that the full model, with all indicators, has a positive mean squared error. Removal of indicators is according to the order of the indicators in the Options Tab, but note that forced indicators are added before free indicators. Removal of indicators is reported in the log window.

As an example suppose that there are 6 indicators with the following linear relations 05 = 01 + 02, and 06 = 03 + 04. The table below list which indicators are aliased in different situations.

Order of indicators	Free	Forced	Aliased
01 02 03 04 05 06	01 02 03 04 05 06		05 06
01 02 03 04 05 06	01 02 03 04	05 06	02 04
01 02 03 04 05 06	01 04 05	02 03 06	04 05
02 05 03 01 06 04	01 02 03 04 06	05	01 04

9.6 Comments on Scenario analysis

Usually, crops have certain optimum conditions under which they thrive well. Too dry or too wet may cause similar yield reductions. The same is true for other variables such as temperature. Scenario analysis aims to identify historical years during which the agro-meteorological conditions were similar to the current year or target year.

The scenario approach is especially meant for predicting the yields during years with exceptional weather. The time trend is supposed to represent the general tendency for "normal" years. It should be realized that there are often certain growth stages during which crops are more sensitive to abnormal conditions than usual. The flowering stage is of course the most obvious example. At times, small changes in the sequence of meteorological events have major effects in crop response.

It is recommended that the selection of indicators for Principal Component Analysis is done based on "non-statistical insights". A nice set of similar years which seem

similar may easily be obtained by clicking a few buttons. However, if those years are not really similar to the target year, one may end up predicting a higher yield than the time trend would suggest whereas the agro-meteorological conditions would rather suggest a lower yield than usual!

After establishing a set of similar years, it is good to check why they are marked as similar. Are there years which are very similar to the target year and others less similar? The score plots can answer this question. Which indicators contribute considerably to the selected factors? The table with the loadings can answer this question. And would there be a causal explanation as to how those indicators could affect crop yield in a favourable or unfavourable way? Maybe such explanations cannot be given for all indicators used in constructing the selected factors.

A found similarity should be regarded as real when many of the included indicators have indeed got similar values for all the so-called similar years. When a causal explanation can be pointed out also as to how the crop yields could be affected, a similarity should be regarded as meaningful. In that case, the similar years are expected on one side of the trend line – i.e. in the plot of yields versus time.

The criterion "residual standard deviation" or RSD - as used for automatically selecting the best model – will be evaluated at a later stage. Possibly a better criterion can be found for this. The presumption that the target year is an exceptional year in one way or another should be regarded as more important than what the RSD values indicate.

In the case of regression analysis, the result from the time trend analysis is taken as indicative but the exact coefficients are fitted again after one or more indicators are selected for the model. In the case of scenario analysis, the results from the time trend analysis are fully adopted and taken as a starting point for the scenario analysis. In that sense, the time trend analysis is not as fully integrated into scenario analysis as is the case with regression analysis. For that reason, one can say that scenario analysis involves applying "a correction" to the time trend, based on the results from the Principal Component and the Cluster Analysis.

We would like to recommend a way to integrate the time trend analysis further into the scenario analysis: an extra variable could be introduced which indicates the similarity of each year. This variable can e.g. be a dummy variable with only zeroes and ones. This variable can then be used as extra parameter in a regression model, together with the constant, possibly linear and maybe even the quadratic term of the time trend.

Figure 8.2 graphically shows what this would mean for a linear time trend. A dummy variable was introduced to indicate whether a year is similar or non-similar (1 and 0 respectively). The yields for similar years are plotted with square markers and non-similar years with triangular ones. The general time trend through all these points is indicated by the dotted line. Due to the introduction of the dummy variable, there

are now two time trends: the dashed line indicates the time trend for the non-similar years and the drawn line the time trend for the similar years.

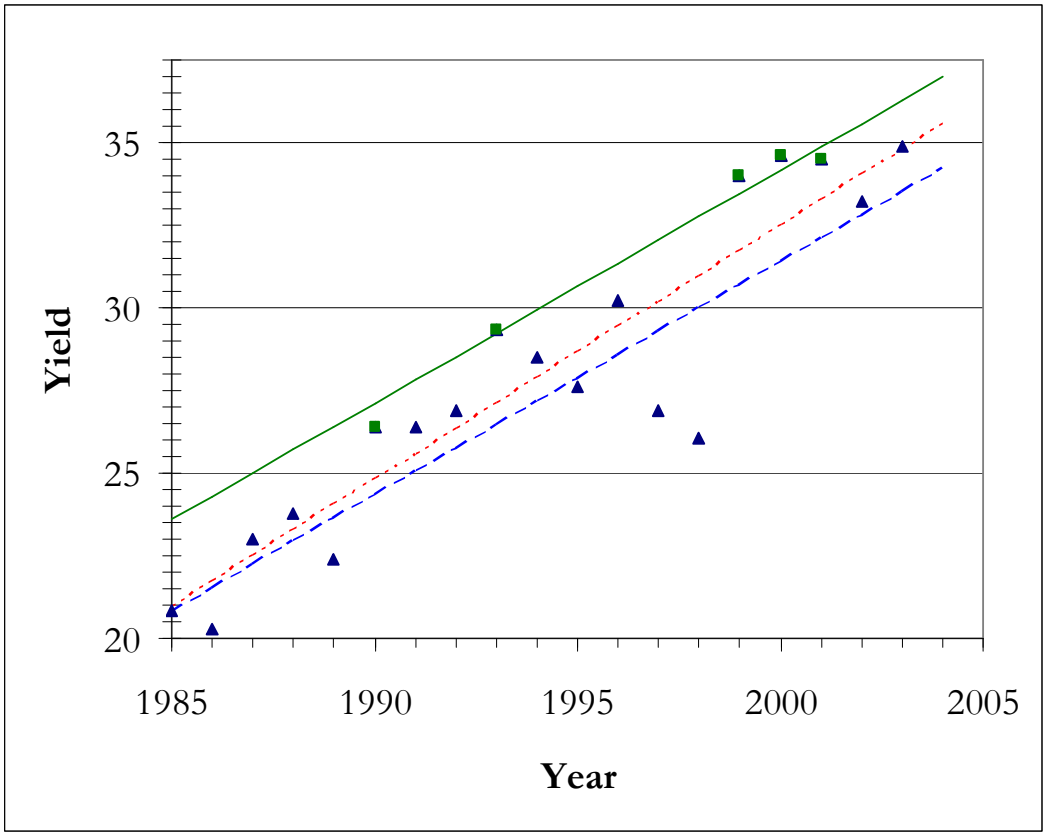


Figure 8.2: model with linear time trend and a dummy variable

Please note that the figures in the dummy vector sum up to N – i.e. the number of similar years. It should be furthermore noted that the slope of the general time trend is not necessarily the same as the slope of the two new time trends. In rare cases, the general time trend might even cross with one of the new time trends within the period under consideration. Therefore, even if the similar years all have higher yields than the non-similar years, the prediction based on the general time trend could in rare cases still be higher than the one based on the time trend for the similar years.

In our opinion, by introducing the dummy variable the results from the scenario analysis are further integrated into the forecasting process than in the current corrective approach. It is also easier to calculate various statistics as well as confidence intervals for the forecasts. This can be done in the same way as is done for the forecasts based on regression analysis. In this way, results from scenario analysis will become more comparable with those obtained by means of regression analysis.

The model could be adapted further by using a vector with weights instead of the dummy variable, so that among the similar years those very close will have more influence than those which are almost as far away as the cutoff distance. The weights could be derived from figures in the distance matrix. The weights should preferably

be scaled, so that they also sum up to N – i.e. the number of similar years. Which method should be chosen for determining the weights seems a difficult question though. This could introduce an element of arbitrariness and could thus make the introduction of a vector with weights also less convincing.

10 Installation, databases and analyst settings

In the following the installation, the database and analyst settings will be described. For the sake of completeness, a short paragraph is added with information about text files used by CgmsStatTool.

10.1 Installation

CgmsStatTool will be installed by default on the following directory: C:\Program Files\Alterra\CgmsStatTool. Under this directory you will find the following sub-directories:

- Doc: includes this user manual
- Images: images supporting the user interface
- Log: application error messages
- Maps: maps to support selection of an area or region. *Note that these can be omitted in case of applying CgmsStatTool in an other region of interest. In such case the drop down list can be used only. Alternatively such maps could be produced at request.*
- Queries: queries used to retrieve data from the database
- Resources: miscellaneous files e.g. statistics.txt which describes the codes for the summary statistics to be displayed of a selected model

On the main directory the following files are present:

- CgmsStatTool.exe: the CgmsStatTool
- DFORRT.DLL: Visual Fortran Runtime library
- StatLevel3.dll: the Fortran DLL responsible for determining the models

Besides files are installed under:
`C:\Documents and Settings\%Username%\Application Data\Alterra\CgmsStatTool\`
in which %Username% is the username of the actual user. These files are:

- Cgmsstattool.csv: saved options, which were used for the analysis (see Annex 3)
- indicatornames.txt: mapping of indicator code and their name (further details can be found in last section of Annex 1)
- CgmsStatTool.ini: configuration options of CgmsStatTool (see Annex 4)

These files are saved in this directory because they are user specific. Upon installation of the tool, the above-mentioned directory is created only for the user who installs the tool. Besides a similar directory is created for all other users under this directory: C:\Documents and Settings\All Users\. This directory also includes the file Settings.mdb: a Microsoft Access database with link to Cgmsstattool.csv for maintenance purposes. When another user runs CgmsStatTool on the same computer, the tool will by default try to work with his (or her) own analyst settings. However, if the right directory is not found, the tool will revert to the directory for

"All Users". If a user wants to work with his (or her) own settings on a computer where CgmsStatTool is already installed, he (or she) can copy the directory named "CgmsStatTool" under "All Users" to the same location in his own directory tree under C:\Documents and Settings\, i.e. including a number of other files such as " CgmsStatTool.ini".

Finally the CgmsStatTool is supplied with a sample Access database, cgms.mdb, which can be found under C:\Documents and Settings\All Users\Application Data\Alterra\data\. The ODBC link called "CGMS local database" that is created during installation refers to this cgms.mdb. See for more information on changing database Annex 2.

10.2 Database and analyst settings

The CgmsStatTool interacts with other tools and programmes via the database. For the mean time, the tool can only work with one database. No facilities have been included as yet to enable the user to switch database. CgmsStatTool is meant to work in the first place with a database created in Microsoft Access. Predicted yields can be saved to this database. CgmsStatTool can also run in batch mode – see chapter 11. When run in that mode, predicted yields are calculated based on so-called analyst settings and written to the Oracle production database. It is expected however that a facility for switching database will be added in the future, for use in the interactive mode.

When a user chooses to save the results of an analysis to the database, the user also saves the options, which were used for the analysis, to a text file at the same time. In case of regression analysis 26 lines are saved and in case of scenario analysis 17 lines. We will refer to these saved options as analyst settings. Please note that in principle, these settings are user specific. The mentioned 26 lines – or 17 line - can easily be extracted from the text file – e.g. by using a simple text editor. When those lines are saved to another text file with extension "csv", they can be shared with other users of CgmsStatTool who will be able to open that text file in CgmsStatTool and they will be able to evaluate the analysis right away, provided their databases contain the same data.

10.2.1 Database

The tool retrieves data from a database and results can be written to that database too. The database containing the yield and indicator data is assumed to contain at least the following tables:

- DATA_FOR_YIELD_FORECAST
- EUROSTAT
- NUTS
- STAT_CROP
- RUN

The structure of these tables is described in further detail in Annex 1. Besides, the following tables are required, if the user wants to save results of his regression analyses to disk:

- FORECASTED_NUTS_YIELD
- MODEL_EXCL_YEARS
- MODEL_INCL_INDICATORS
- MODEL_REGR_INDICATIFS

The structure of these tables is also described in further detail in Annex 1. The last three tables are new ones. The table MODEL_REGR_INDICATIFS is the most important one of the three. This table stores the selected model almost completely:

1. The selected start year;
2. The selected end year;
3. The offset used for the years;
4. The selected transformation for the years;
5. The selected time trend;
6. Coefficient of selected time trend;
7. A number of summary statistics for the selected model.

In addition, the tables MODEL_EXCL_YEARS and MODEL_REGR_ INCL_INDICATORS store:

1. The years which were excluded; and
2. The indicators which were included in the model and their coefficients

If the user wants to save results of his scenario analyses to disk, the following additional tables are required:

- MODEL_SCEN_INDICATIFS
- MODEL_SCEN_SIM_YEARS

CgmsStatTool is supplied with a sample Access database, containing all the mentioned tables. CgmsStatTool was created with the Microsoft Access database platform as first target in mind, to be accessed via ODBC. However, CgmsStatTool can in principle make use of other database management systems as well such as Oracle, SQL Server, Interbase or Firebird – although no other database management system than Microsoft Access has been tested to work with the tool. With some database management systems, the tool could even access an alternative database without making use of ODBC. Annex 2 describes how the tool can be made to work with another database.

10.2.2 Analyst settings

In the analyst settings, other things are stored than what is stored in the database when the user presses "Save". A complete set of analyst settings, spanning either 17 or 26 lines, is specific for a specific area, crop, decade and analysis type. The model

that has led to the prediction is not stored as such, but the options selected by the analyst are stored which led to that model, e.g. which indicators are included as free and which ones are included as forced indicators. Hence the analysis can be reviewed at any later time and if necessary it can be repeated for changed circumstances or even for the year following the one for which the analysis was carried out originally.

All analyst settings are stored together with the NUTS code, crop number, decade and analysis type to which they apply. To be more exact: a line in the settings file always starts with four fields separated by a comma, indicating NUTS code, crop number, decade and analysis type respectively. The file format is called comma separated values file or CSV.

Annex 3 describes the analyst settings in further detail as well as the way they are stored, the way they can be copied, the way they can be manipulated etc. It should be noted that the CSV format for storing these settings was chosen because it can be accessed by many programs in different ways, thus giving the user ample freedom to edit, copy, modify or share settings. When editing the file, the user however needs to make sure that the first five fields of every line are unique – i.e. NUTS code, crop number, decade number, analysis type and setting name. Otherwise, an error message "Duplicate index value. Operation aborted" will appear, when the user attempts to open the CSV file in CgmsStatTool.

With analyst settings, a user can retrieve selections he (or she) has made earlier for one particular area, crop, decade and analysis type. However, a user may like to choose his (or her) own general defaults for the options in the interface of CgmsStatTool. Annex 4 describes how these defaults are stored in the INI file of CgmsStatTool and how these defaults can be manipulated.

10.3 Text files

As mentioned, CgmsStatTool retrieves data from a database and from a file with analyst settings. Besides, the tool also retrieves indicator names and names of summary statistics from the following text files respectively:

1. Indicatornames.txt
2. Statistics.txt

The file "indicatornames.txt" may need updating, viz. when new indicators become available in the database. Further details can be found in last section of Annex 1.

The file "statistics.txt" describes the codes for the summary statistics to be displayed of a selected model.

11 Menu Items and Clickable Icons

11.1 Interactive use

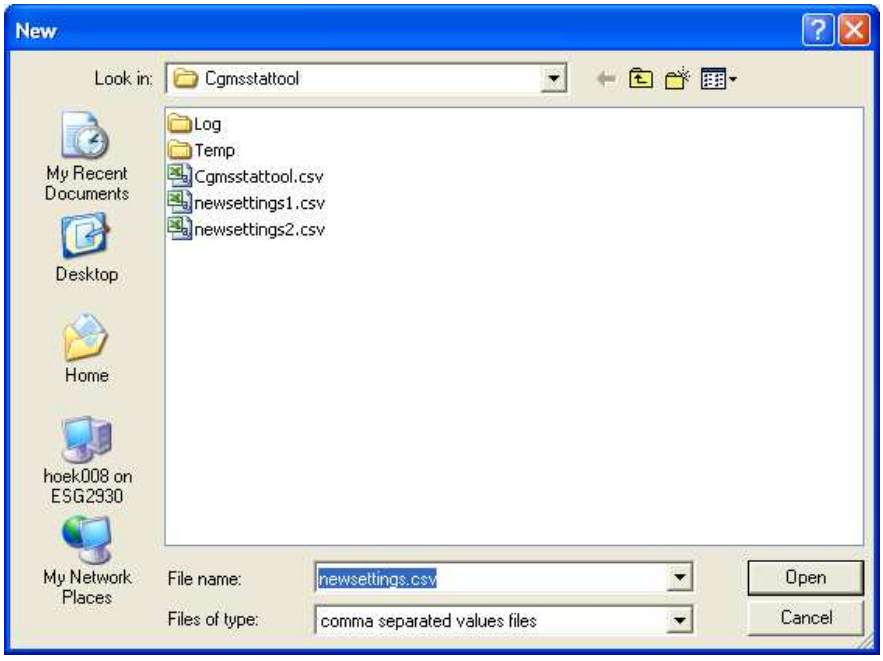
In most cases the tool can be driven by selecting options offered in the left menu and options offered on one of the five tab pages on the right panel. However, the tool is equipped with a menu bar as well as clickable items for special actions.

The menu has the following structure:

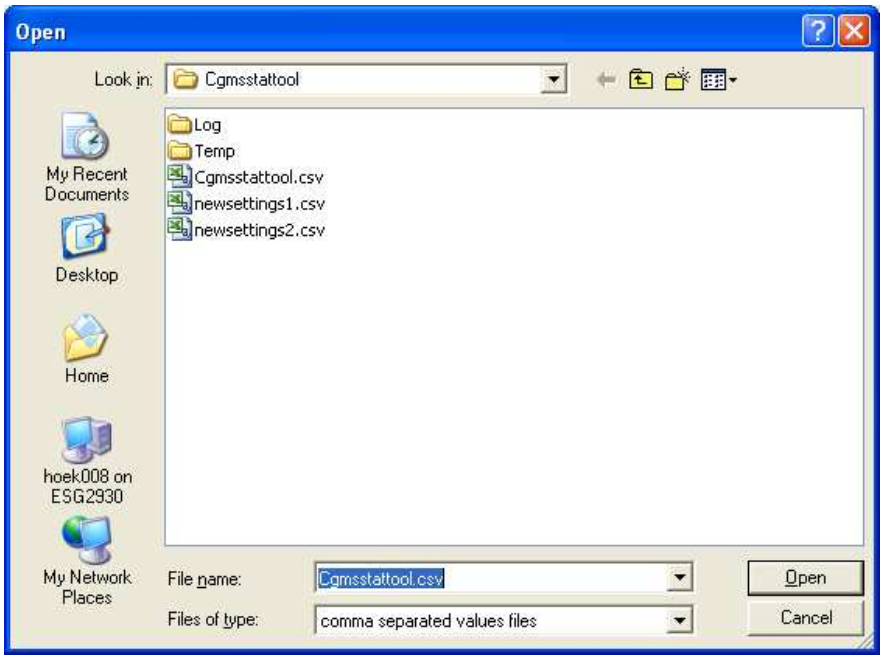
- File New; Open; Save; Save As; Print; Exit;
- View Log Window;
- Tools Run; Connections;
- Help About.

The bar with clickable icons looks like this:

At start up, the programme by default tries to retrieve the file "CgmsStatTool.csv" with analyst settings from a user directory – i.e. from a user specific directory under "C:\Documents and Settings\" if the user is running Windows 2000 or higher. If the file "CgmsStatTool.csv" does not exist, it is created. However, the user may want to load a different file with analyst settings. The file menu therefore has the options "New" and "Open". When "New" is selected, an empty CSV file is created.



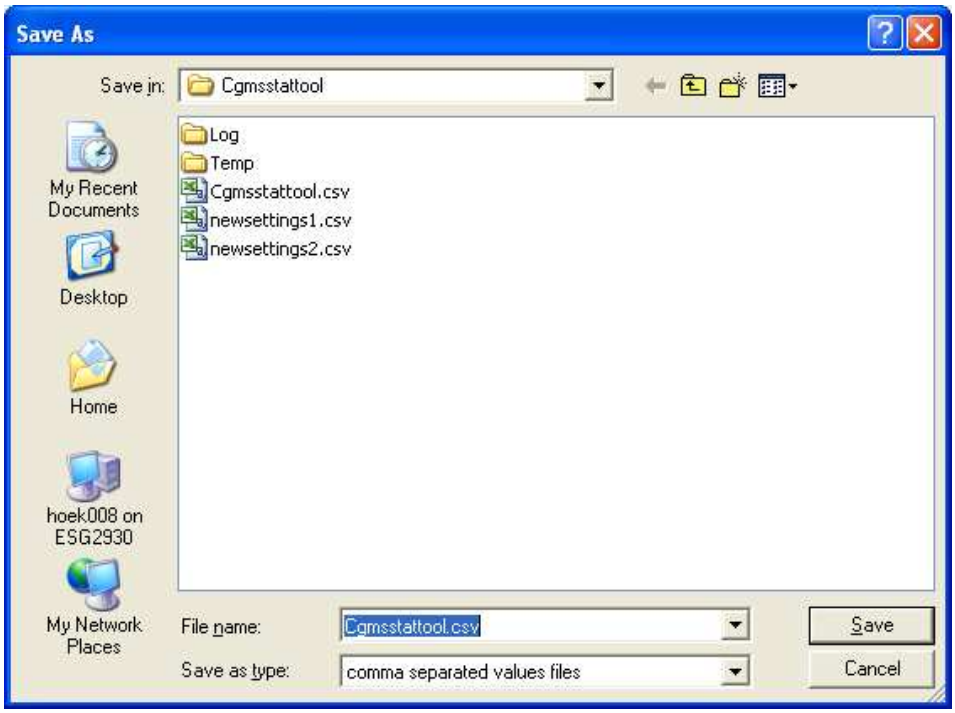
When the option "Open" is selected, the user is offered the chance to select another file to work with.



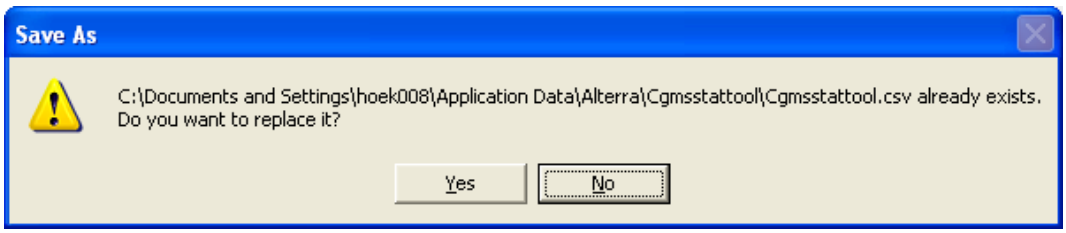
The "Open" dialog always suggests the name of the currently loaded file. The same dialog appears when the open button is pressed, i.e. the first button on the icon bar. After selecting another file, a message appears in the log window. Of course the programme continues to work with the indicated file from then on, until it is terminated. If the user wants to work with a new file different from "CgmsStatTool.csv", the user can create a new text file in the Windows explorer and

change the extension to "csv". The user can then open the newly created file and start to fill that file with a new set of analyst settings.

Adding analyst settings to the currently loaded CSV file is done by choosing the "Save" option from the File menu. Pressing the "Save" button – the second button on the icon bar has the same effect and the "Save" button on the Output page has that effect too. When any of these buttons is pressed, the currently selected regression model is also saved to the database (MODEL_EXCL_YEARS, MODEL_INCL_INDICATORS and MODEL_REGR_INDICATIFS). At times, the user may like to save the currently loaded analyst settings to another file. In that case the user can choose the "Save As" option from the File menu. A dialog then appears:



The dialog will suggest a file name, which is always the name of the currently loaded file. If the user presses the "Save" button in the dialog without having changed the file name, another dialog appears:



If the user presses "Yes", the analyst settings are saved to the indicated CSV file. If the user presses the "Save" button in the dialog after typing a new file name, the

analyst settings are saved to the indicated CSV file. Please note that the currently selected regression model is not saved to the database when option "Save As" is chosen.

Any time the Output or the Model details page is open, it is possible to print whatever is shown there. The user can start printing to the default printer by choosing "Print" from the File menu or by pressing the "Print" button – i.e. the last button on the icon bar. In the case of the Model details page, a print dialog appears first, in the case of the Output page it does not. In case the user wants to represent the information from the Output page in a way different from the way it comes out from the default printer, the button "Copy to clipboard" at the bottom of the Output page can be of help.

When the option "Exit" is chosen, a dialog pops up asking for a confirmation that you really want to exit CGMSSTATTOOL:

The only option under the "View menu" i.e. "Log Window" – can be used to show or hide the fourth optional panel with a log window which can contain informational messages, warnings and error messages. The log window always appears below the left and right panels. It can be right clicked after which a popup menu appears with the options Hide, Clear and Copy. If the latter option is chosen, always all the messages in the log window are copied to the clipboard. If only a few lines from the log window need to be copied to the clipboard, this can be done by selecting those lines followed by key combination Control C.

The option "Run" under the "Tools" menu can be chosen when a combination of area, crop and period is selected for which yield and indicator data are available. This option is only activated when at least one indicator is selected on the Indicators page. In principle it is always possible though to build a regression model based on the time trend alone. The "Run" button – the third button – on the icon bar does the same as the "Run" option under the "Tools menu". The option "Connections" under the "Tools menu" is not yet implemented.

11.2 Batch mode

Start in batch mode with: CgmsStatTool /batch <filename>

This filename should have the same structure as CgmsStatTool.csv (see section 10.2.2) . The program searches for the following locations in the this order:

- directory where CgmsStatTool.exe is located
- using the path that can be given in <filename>
- using the PATH defined by Windows environment variable
- UserAppDataFolder, e.g.: `C:\Documents and Settings\%Username%\Application Data\Alterra\CgmsStatTool`

A (batch)window is started. In this window the progress of the batch job is shown:

Example:

```
Started at: 19-5-2008 17:16:10
Run batchmode from: C:\Documents and Settings\rolle001\Application
Data\Alterra\Cgmsstattool\asemars_one.csv
Number of settings read: 4
- [25.00%] ForecastRun( FNutsCode=DE, FStatCropNo=4, FTargetYear=2005, FDecade=13,
FRunId=DE04200513R )
- [50.00%] ForecastRun( FNutsCode=DE, FStatCropNo=1, FTargetYear=2005, FDecade=13,
FRunId=DE01200513R )
- [75.00%] ForecastRun( FNutsCode=DE, FStatCropNo=4, FTargetYear=2005, FDecade=14,
FRunId=DE04200514R )
- [100.00%] ForecastRun( FNutsCode=DE, FStatCropNo=1, FTargetYear=2005,
FDecade=14, FRunId=DE01200514R )
Finished at: 19-5-2008 17:16:11
Duration   : 00:00:01
```

This info is also stored in the file CgmsStatTool_batch.log. Results of best forecasts are stored in the database. The criteria that determines the best model, is defined by the setting variable called "BestModelSelection". On forehand models are excluded:

- for which the variance inflation factor exceeds the setting variable "MaxVifMeasure"
- that are wrongly correlated in case the setting variable "HighlightWrongSign" is activated (= -1)
- that are not significant in case the setting variable "HighlightIfNotSignificant" is activated (= -1)

The minimum number of complete years (all selected indicators have a value and yields are available as well) allowing the calculation of a regression model is given in the file CgmsStatTool.ini:

[batch settings]
MinNumAvailYears =6

12 Validation of the CGMS Statistical Toolbox

Curnel and Oger (2006) thoroughly tested and validated a beta version of CgmsStatTool, which contained only features relevant for regression analysis. Most of the problems they encountered were fixed in the final version of CgmsStatTool. Some of the problems they reported were in fact features of the programme. They also wrote down several suggestions as to how to improve the user friendliness of the interface and most of these were implemented in the final version. A detailed reaction to their comments was reported to JRC.

The output of the Fortran subroutines TIMETREND, CORRELATION, FITMODEL, SINGLE and BESTSUBSET was compared with the output of the statistical program GenStat (2005). This was done for a variety of datasets, including datasets with:

- Aliased indicators;
- Indicators that were aliased with the linear time trend;
- Heavily correlated indicators;
- Some free or forced indicators that gave a perfect fit with zero mean square for residual;
- One degree of freedom for residual;
- No indicators but only a time trend;
- Constant yield or yields that were perfectly linear or quadratic in time

In all cases the difference in output between CgmsStatTool and GenStat was negligible and the Fortran programs are therefore firmly validated. The same datasets were used to confirm that the CgmsStatTool interface correctly presents the regression results to the user.

The features relevant for scenario analysis were validated by our colleague De Wit.

13 References

Cook, R.D and Weisberg, S. (1982). Residuals and Influence in Regression. Chapman and Hall. London.

Curnel, Y. and Oger, R. (2006). ASEMARS Lot I Task 4: Improvement of the statistical module of CGMS: Test & validation (subtask a, b & c). Report, Biometry, data management and agrometeorology unit, Walloon Agricultural Centre, Gembloux, Belgium.

Flack, V.F. and Chang, P.C. (1987). Frequency of selecting noise variables in subset regression analysis: a simulation study. The American Statistician, 41, 84 86.

Furnival, G.M. and Wilson, R.W. (1974). Regression by leaps and bounds. Technometrics, 16, 499 511.

GenStat (2005) version 8.2. Statistical Computer Programme. VSN International Ltd. Hemel Hempstead. United Kingdom

Goedhart, P.W. (2005). GenStat procedure RSELECT. In: P.W. Goedhart & J.T.N.M. Thissen (eds.), Biometris GenStat Procedure Library Manual 8th Edition (pp. 85 88). Report, Biometris ,Wageningen UR, The Netherlands.

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models, second edition. Chapman and Hall. London.

Montgomery, D.C., Peck, E.A. and Vining, G.G. (2001). Introduction to Linear Regression Analysis, third edition. Wiley, New York.

Ter Braak, C.J.F. and Groeneveld, A. (1982). SUBSEL een Fortran programma voor "SUBset SELection" in regressiemodellen gebaseerd op subroutines van Furnival en Wilson. IWIS rapport B 82 ST 79 41. Wageningen. The Netherlands.

Ward, J. H. (1963), Hierarchical Grouping to optimize an objective function. Journal of American Statistical Association, 58(301), 236-244.

Annex 1 Structure and content of the database

The following scripts could be used to create the tables in Microsoft Access:

```
CREATE TABLE DATA_FOR_YIELD_FORECAST (
  NUTS_CODE VARCHAR(8) NOT NULL,
  YEAR DOUBLE NOT NULL,
  DECADE DOUBLE NOT NULL,
  STAT_CROP_NO DOUBLE NOT NULL,
  INDICATOR_CODE VARCHAR(50) NOT NULL,
  INDICATOR_VALUE DOUBLE
);

CREATE TABLE EUROSTAT (
  STAT_CROP_NO DOUBLE NOT NULL,
  NUTS_CODE VARCHAR(8) NOT NULL,
  YEAR DOUBLE NOT NULL,
  AREA_CULTIVATED DOUBLE NOT NULL,
  OFFICIAL_YIELD DOUBLE
);

CREATE TABLE NUTS (
  NUTS_CODE VARCHAR(8) NOT NULL,
  NUTS_NAME VARCHAR(60),
  NUTS_LEVEL DOUBLE,
  BELONGS_TO VARCHAR(8),
  USED_CGMS VARCHAR(1),
  LEVEL_CGMS DOUBLE
);

CREATE TABLE STAT_CROP (
  STAT_CROP_NO DOUBLE NOT NULL,
  STAT_CROP_NAME VARCHAR(40) NOT NULL,
  SIM_CROP_NO DOUBLE NOT NULL,
  SIM_CROP_NAME VARCHAR(40) NOT NULL
);
```

The tables needed to store the results of CGMSSTATTOOL could be created with these lines:

```
CREATE TABLE RUN (
  RUN_ID VARCHAR(25) NOT NULL,
  NUTS_CODE VARCHAR(8) NOT NULL,
  STAT_CROP_NO INTEGER NOT NULL,
  TARGET_YEAR INTEGER NOT NULL,
  DECADE INTEGER NOT NULL,
  ANALYSIS VARCHAR(1) NOT NULL,
```

```
  RUN_CODE VARCHAR(25) NOT NULL
)

CREATE TABLE FORECASTED_NUTS_YIELD (
  RUN_ID VARCHAR(25) NOT NULL,
  FORECASTED_YIELD DOUBLE
);

CREATE TABLE MODEL_EXCL_YEARS (
   RUN_ID VARCHAR(25) NOT NULL,
   EXCLUDED_YEAR INTEGER NOT NULL
);

CREATE TABLE MODEL _INCL_INDICATORS (
  RUN_ID VARCHAR(25) NOT NULL,
  INDICATOR_NAME VARCHAR(50) NOT NULL,
  MODEL_COEFF DOUBLE NOT NULL
);

CREATE TABLE MODEL_REGR_INDICATIFS (
  RUN_ID VARCHAR(25) NOT NULL,
  START_YEAR INTEGER NOT NULL,
  END_YEAR INTEGER NOT NULL,
  NUMBER_OF_YEARS INTEGER NOT NULL,
  TRANSFORM_YEARS VARCHAR(12) NOT NULL,
  YEAR_OFFSET DOUBLE NOT NULL,
  ORDER_OF_TIMETREND INTEGER NOT NULL,
  RSQ DOUBLE NOT NULL,
  RSQ_ADJ DOUBLE NOT NULL,
  RES_STD_DEV DOUBLE NOT NULL,
  RMSQ_ERR_PRED DOUBLE NOT NULL,
  MALLOWS_CP DOUBLE NOT NULL,
  MAX_VAR_INFL_FACT DOUBLE NOT NULL,
  STD_ERR_PRED_MEAN DOUBLE NOT NULL,
  STD_ERR_PRED_NEW DOUBLE NOT NULL,
  CONST_COEFF DOUBLE,
  TIMETREND_LIN DOUBLE,
  TIME_TREND_QUAD  DOUBLE
);

CREATE TABLE MODEL_SCEN_SIM_YEARS (
  RUN_ID VARCHAR(25) NOT NULL,
  SIMILAR_YEAR INTEGER NOT NULL,
  EUCLID_DISTANCE DOUBLE
)
```

```
CREATE TABLE MODEL_SCEN_INDICATIFS (
  RUN_ID VARCHAR(25) NOT NULL,
  START_YEAR INTEGER NOT NULL,
  END_YEAR INTEGER NOT NULL,
  NUMBER_OF_YEARS INTEGER NOT NULL,
  TRANSFORM_YEARS VARCHAR(12) NOT NULL,
  YEAR_OFFSET DOUBLE NOT NULL,
  ORDER_OF_TIMETREND INTEGER NOT NULL,
  PCOMP_COUNT INTEGER,
  PERC_EXPL_VARIANCE DOUBLE,
  CUTOFFD DOUBLE,
  SCEN_MODEL_TYPE VARCHAR(8),
  RES_STD_DEV DOUBLE NOT NULL,
  CONST_COEFF DOUBLE,
  TIMETREND_LIN DOUBLE,
  TIME_TREND_QUAD  DOUBLE
)
```

The content of the table DATA_FOR_YIELD_FORECAST will be updated more often than any of the other tables. New indicators may be added from time to time. In that case, the text file "indicatornames.txt" (see section 10.3) may have to be updated, for the tool to work properly with the database. The need may arise for extra lines in that text file, to make sure that the indicator codes in the database can be matched with a user-friendly description. At the moment, a maximum of 30 indicator code name mappings can be included in this file. The names need to start with a two-digit number (01 thru 30).

Annex 2 How to configure the tool for another database

To understand how CgmsStatTool can be configured to use another database management system, one will need to understand how the tool works with Microsoft Access. When the tool is installed, an ODBC link is created called "CGMS local database". As part of the installation, a so-called INI file is provided (Asemars.ini), which contains a reference to this ODBC link. This reference is read from the INI file when the tool is started and it is then used to make the connection to the database.

For the tool to work with an alternative database, an ODBC driver will need to be available for that database management system, or otherwise a BDE driver. Of course, a database instance will need to created, the mentioned tables will need to be created therein and those will have to be filled with data. Furthermore there will be a need to make the functions LCASE and SGN available, which are specific for Microsoft Access. In ORACLE these function are as follows:

```
CREATE FUNCTION LCASE (ASTR VARCHAR(60))
AS
BEGIN
  RETURN LOWER(ASTR)
END

CREATE FUNCTION SGN (ANUM DOUBLE)
AS
BEGIN
  RETURN SIGN(ANUM)
END
```

For connecting to an Oracle database instance, the following two drivers are recommended: Oracle ODBC Driver from Oracle and the BDE ORACLE driver from Borland. We have tried to run the tool also with the Microsoft ODBC for Oracle driver, but we experienced too many errors. The scripts described in Annex 1 could help to make the tool work with an Oracle database. Note that tables DATA_FOR_YIELD_FORECAST, EUROSTAT, NUTS and STAT_CROP are assumed to be available in the Oracle database already.

To make the tool work with an alternative database, one will also need to create an ODBC link, open the INI file "CgmsStatTool.ini" and replace the name of the existing ODBC link there with the name of the newly created ODBC link. In case one decides to use a BDE driver, a BDE alias has to be created and the name of that alias will need to be entered into the INI file instead.

A list of existing ODBC links can be obtained by opening the ODBC Data Source Administrator. On Windows XP, this tool can be found in the Start Menu under Settings - Control Panel - Administrative Tools. New ODBC links can also be

created here. The ODBC link called CGMS Local Database used by CGMSSTATTOOL can be found on the tab sheet "System DSN". The implication of this is that the ODBC link can be used by all users of the computer. If two or more users of a computer want CgmsStatTool to use different databases, then each of them should create a personal ODBC link on the tab sheet "User DSN".

In addition, the text file "indicatornames.txt" may have to be updated, for the tool to work properly with the alternative database. The final section of Annex 1 contains more information on how to update this text file.

Annex 3 Analyst settings

Analyst settings are by default saved in file called "CgmsStatTool.csv" in the following directory:

```
C:\Documents and Settings\%Username%\Application Data\Alterra\CgmsStatTool\
```

in which %Username% is the username of the actual user. They are saved in this directory because they are user specific. Upon installation of the tool, the above-mentioned directory is created only for the user who installs the tool. Besides a similar directory is created for all other users under this directory:

```
C:\Documents and Settings\All Users\
```

When another user runs CgmsStatTool on the same computer, the tool will by default try to work with his (or her) own analyst settings. However, if the right directory is not found, the tool will revert to the directory for "All Users". If a user wants to work with his (or her) own settings on a computer where CgmsStatTool is already installed, he (or she) can copy the directory named "CgmsStatTool" under "All Users" to the same location in his own directory tree under C:\Documents and Settings\, i.e. including a number of other files such as " CgmsStatTool.ini".

The user has the freedom to save settings to another CSV-file in another directory. Such CSV-files with settings can be opened at any time from the menu: File - Open. Alternatively, the user can press the Open button.

The files containing the analyst settings need to follow a special CSV format – which we'll call CgmsStatTool CSV format. This format is described further below. In the following it is first explained how to use these settings files. The CgmsStatTool CSV format can be read and written by CgmsStatTool, but besides it can be opened and manipulated by many other programs, such as:

- Text editors
- Microsoft Excel and similar spreadsheet programs – viz. by importing the CSV-file, not by opening it!
- Microsoft Access – i.e. by creating a linked table referencing the CSV-file

Of course, care should be taken when manipulating the settings file, in order not to lose previously entered settings. It is strongly recommended to always backup a copy of such a settings file before manipulating it. The user particularly needs to make sure that the first five fields of every line are unique – i.e. NUTS code, crop number, dekad number, analysis type and setting name.

Probably the safest and most powerful way to manipulate such a settings file is by creating a linked table in Microsoft Access, after which the settings can be retrieved, updated and appended using the well-known Structured Query Language (SQL). One could even register the Access database file in the ODBC Data Source Administrator – or in other words create an ODBC link - after which it can be accessed by other database manipulation programs as well as scripts - as long as they work with SQL.

All analyst settings are stored together with the NUTS code, crop number, decade and analysis type to which they apply. To be more exact: a line in the settings file, always starts with four fields separated by a comma, indicating NUTS code, crop number, decade and analysis type. The following line contains an example: "BE","1","28","R","StartYear","1995". This line applies to Belgium, crop number 1 (wheat), dekad 28 (October I) and regression analysis. The actual setting is called StartYear and in this case it has value 1995. If the CSV-file has not been corrupted, there should be 26 other lines found starting with "BE","1","28","R" together with this line. Note the CgmsStatTool is case sensitive for the values read from this file.

The following table gives an overview of which analyst settings are saved in case of regression analysis:

Analyst setting	**Purpose**
StartYear	the first year used for fitting the time trend
EndYear	the last year used for fitting the time trend
TargetYear	the year for which the yield should be predicted
ExcludedYears	which years in the interval StartYear .. EndYear have been excluded; the value should be a comma-separated string with one or more years
TrendModelType	the five trend model that can be selected: None, Linear, Quadratic, AutoUpToLinear (=automatic testing up to linear), AutoUpToQuadratic (=automatic testing up to quadratic)
TimetrendSignificanceLevel	how high should the p-value be before a time trend is considered significant
TransformType	whether the years should be transformed before an attempt is made to fit a time trend; valid values are None and Logarithmic
YearOffset	before a possible transformation is carried out and a time trend is fit, an offset is always subtracted in order to make sure the fitted coefficient(s) do not become very small figures
OrderOfTimetrend	indicates whether there's a time trend or not (0) and if so whether that trend is linear (1) or even quadratic (2)
NumInclYears	number of included years; in principle, this is a redundant value because the number can be calculated from the StartYear, the EndYear and from the ExcludedYears; it is added for convenience
FreeIndicators	indicates which indicators have been selected as free candidates for the regression models; the value should be a comma-separated string with one or more indicator codes
ForcedIndicators	indicates which indicators have been forcibly included into the regression models; the value should be a comma-separated string with one or more indicator codes
ModelingMethod	indicates the method used for generating the regression models; valid values are SingleFree and BestSubset

ModelOrdering	indicates which summary statistic should be used to order the generated regression models; valid value is a figure from the following set: {3, 4, 5, 6, 8, 10, 11} 3 = R-squared 4 = R-squared adjusted 5 = Mallows Cp 6 = Residual standard deviation 8 = Root mean squared error for prediction 10 = Standard error of prediction for mean 11 = Standard error of prediction
BestModelSelection	indicates which summary statistic should be used to select the best model from the generated regression models; as at now, the value for this setting is always the same as the one for user setting ModelOrdering
HighlightWrongSign	specifies whether the t-values of the indicators should be highlighted on the output page if they have the wrong sign (highlighted = -1, not highlighted = 0)
HighlightIfNotSignificant	specifies whether the t-values of the terms of the model should be highlighted on the output page if they are not significant (highlighted = -1, not highlighted = 0)
HighlightSignificanceLevel	specifies the level of significance that is considered just enough for a term to be meaningful
StatsToDisplay	indicates which four summary statistics from the file "statistics.txt" should be displayed on the output page; the line contains a comma-separated string with the codes for the summary statistics to be displayed
MaxVifMeasure	specifies the maximum variance inflation factor beyond which a term should be considered too closely correlated with one of the other terms
MaxNumFreeIndicators	maximum number of free indicators in a model
MaxNumModelsInSubset	maximum number of models to be displayed from a subset
SignsOfIndicators	indicates the desired sign for each of the n indicators in use; a valid value is a string of length n with a "0" for no specified sign at position x for the x'th indicator, a "-" for a negative and a "+" for a positive sign
RanksOfIndicators	indicates the desired rank for each of the n indicators in use; a valid value is a string of length n with a number or letter at position x for the x'th indicator; ranking starts with a "0" for the most important indicator until "9" and afterwards continues with A, B etc.
NumberOfFreeIndicators	number of free indicators in a model (should correspond to the number of indicators mentioned in setting FreeIndicators)
NumberOfForcedIndicators	number of forced indicators in a model (should correspond to the number of indicators mentioned in setting ForcedIndicators)

In case of scenario analysis, the same first ten settings are saved: StartYear through NumInclYears. The following table gives an overview of which other analyst settings are saved in this case:

Analyst setting	**Purpose**
IncludedIndicators	indicates which indicators have been selected as for the Principal Component Analysis; the value should be a comma-separated string with one or more indicator codes
MinPCompCount	minimum number of Principal Components
MinExplVariance	minimum level of variance which is explained by the components
CutOffD1	first cutoff score
CutOffD2	second cutoff score
MinSimYears	minimum number of similar years
MinObs	minimum number of observations; also needed are observations pertaining to the target year.

Annex 4 Configuration options

The following table shows which configuration options are available in the INI file for customizing defaults for the options available in the program interface.

Setting	Purpose and possible values
Section: Database settings	
DSN	data source name, or ODBC link pointing to the database
Username	username required for accessing the database (in case of Microsoft Access this is not needed)
Password	password required for accessing the database (in case of Microsoft Access this is not needed)
Dbms	database management system, e.g. Access or Oracle; at the moment this setting is not essential, but it may become more important if the tool is made to work with DSN-less connections in the future
SettingsFileExt	applies to the extension associated with the format that is selected for storing the user specific settings; at the moment the only valid value is "csv".
Fields	this line describes the fieldnames used in the CSV file; please do not edit this line
FieldSizes	this line describes the sizes of the fields in the CSV file; it is a comma-separated string with field lengths; please do not edit this line
FieldTypes	this line describes the type of the fields in the CSV file; it is a comma-separated string with field types; please do not edit this line
IndexName	to speed up search functionality, the kbmMemTable used in the programme should have an index; this is the name used internally for that index; please do not edit this line
IndexType	for CgmsStatTool, this must be set to Unique, but in theory the kbmMemTable component allows the values Descending, CaseInsensitive and / or NonMaintained; please do not edit this line
IndexDef	this line specifies which fields are used to define the index; it is a semicolon-separated string with fieldnames; please do not edit this line

Section: Default interface settings	
ShowLogWindow	specifies whether or not the log window at the bottom of the tool interface should be shown at startup (False of True)
DefaultSignificanceLevel	specifies the default significance level used in particular in the TimeTrend frame
CorrelationFilterValue	specifies the lowest correlation that should by default be shown in the correlation form; the value should of course be

	in the range 0.0 to 1.0
ModelingMethod	specifies whether the SingleFree method should be used by default or rather the BestSubset method
BestModelSelection	specifies which summary statistic should be used by default to select the best models; the same statistic will be used to rank the various models
HighlightWrongSign	specifies whether the t-values of the indicators should by default be highlighted if they have the wrong sign
HighlightAtSignificance	specifies whether the t-values of the terms of the model should by default be highlighted if they are not significant
IndicatorSignificance	specifies the level of significance that by default is considered just enough for a term to be meaningful
DisplayBasedOnVif	specifies whether models should by default be filtered based on variance inflation factor
MinNumPComps	minimum number of Principal Components
MinPercOfVar	minimum level of variance which is explained by the components
CutoffScore1	first cutoff score
CutoffScore2	second cutoff score
MinNumSimYears	minimum number of similar years
MinN	minimum number of observations; also needed are observations pertaining to the target year.

Section: Default model settings	
TrendModelType	specifies the default trend model; valid values are: None, Linear, Quadratic, AutoUpToLinear, AutoUpToQuadratic
TransformYear	specifies the default choice whether the years should be transformed; valid values are None and Logarithmic
YearOffset	specifies the default offset that is subtracted from the year in case of a logarithmic transformation

Section: Default output settings	
ShrinkingWindowMinimumSize	default minimum number of years for fitting a time trend
StatsToDisplay	indicates which four summary statistics from the file "statistics.txt" should by default be displayed on the output frame; the line contains a letter for each of the 10 statistics with the letter Y at position x meaning that the x-th statistic should be displayed and the letter N meaning that it should not
MaxVifMeasure	specifies the default maximum variance inflation factor beyond which a term should be considered too closely correlated with one of the other terms
MaxNumFreeIndicators	default maximum number of free indicators in a model

MaxNumModelsInSubset	default maximum number of models to be displayed from a subset
DebugLevel	parameter which can be varied between 0 and 2 in order to get less or rather more informational messages, warnings and error messages
MaxWeight	weights used for calculating a correction to the time trend in the scenario analysis should not exceed this maximum

Section: Batch settings	
MinNumAvailYears	Minimum number of years, for which indicator values and yields are available, for calculating regression models.
SecToCloseBatchWindow	Seconds before closing the log batch window

Annex 5 Acronyms and abbreviations

Agrifish	A unit of the EU Institute for the Protection and Security of the Citizen (IPSC) which is part of the Joint Research Centre (JRC) and which provides technical support to DG Agriculture and Member-States (area based subsidies management and control, EU crop yield forecasts for CAP decision making), to DG Regional Policy and DG Enlargement (Land Administration), to DG Fisheries and Maritime Affairs (vessel monitoring, scientific advice) and EU aid & assistance policies (Food security assessment). Services rely on remote sensing (satellite and aerial imagery), agro-meteo modelling, survey, geomatics and GIS and IT techniques.
Alterra	Alterra is the research institute of Wageningen UR concerned with our green living environment. Alterra offers a combination of practical and scientific research in a multitude of disciplines related to the green world around us and the sustainable use of our living environment. Flora and fauna, soil, water, the environment, geo-information and remote sensing, landscape and spatial planning, man and society: these are a few of the numerous aspects of our green environment that Alterra focuses on.
Asemars	Actions in Support of the Enlargement of the MARS Crop Yield Forecasting System. The purpose of this project is - among other things - to complete and reinforce the current version of CGMS in order to extend the system thematically and geographically.
Biometris	Department of Wageningen UR specialised in Applied Statistics, involved in research as well as education
C++	A powerful, general-purpose, high-level programming language with low-level facilities which has been in use since 1985. It is a multi-paradigm language supporting procedural programming, data abstraction, object-oriented programming and generic programming. It was made an ANSI standard in 1998. Many operating systems were developed using C++ including Windows and Linux. Programmes written in C++ for a specific operating system, can often be ported easily to other platforms.
CGMS	Crop Growth Monitoring System, provides the European Commission (DG Agriculture) with objective, timely and quantitative yield forecasts at regional and national scale. CGMS monitors crops development in Europe, driven by meteorological conditions modified by soil characteristics and crop parameters. This mechanistic approach describes crop cycle – e.g. biomass in combination with phenological development from sowing to maturity on a daily time scale. The main characteristic of CGMS lies in its spatialisation component, integrating interpolated meteorological data, soils and crops parameters, through elementary mapping units used for simulation in the crop model.
CSV	Comma Separated Values format, is a delimited data format that has

	fields separated by the comma character and records separated by newlines. Fields that contain a comma, newline, or double quote character, or which start or end with whitespace that is to be preserved, must be enclosed in double quotes. However, if a line contains a single entry which is the empty string, it may be enclosed in double quotes. If a field's value contains a double quote character it is escaped by placing another double quote character next to it. The CSV file format does not require a specific character encoding, byte order, or line terminator format.
DFFITS	A statistic which is a scaled measure of the change in the predicted value for the i-th observation. Large absolute values of this statistic for a certain i indicate that thi i-th observation is influential.
Delphi	A software development package created by Borland Software Corporation. It was first published in 1995 as one of the first Rapid Application Development tools for the Windows operating system. From the beginning, the Delphi development environment supported a special variant of Object Pascal, also known as the Delphi programming language.
DG Agriculture	The European Commission's Directorate-General for Agriculture and Rural Development is based in Brussels. With a staff of about 1000 it is responsible for the implementation of agriculture and rural development policy, the latter being managed in conjunction with the other DGs which deal with structural policies. It is made up of twelve Directorates dealing with all aspects of the Common Agricultural Policy (CAP) including market measures, rural development policy, financial matters as well as international relations relating to agriculture.
DG Eurostat	Directorate General Eurostat, Statistical Information Service of the EU
DLL	Dynamic Link Library, a computer library that implements the concept of dynamic linking. This term is often shortened to DLL. In Microsoft Windows, linking to dynamic libraries is usually handled by linking to an import library when building or linking to create an executable file.
EUROSTAT	Statistical Information Service of the EU
Fortran	One of the first programming languages, first developed by IBM in the 1950s for scientific and engineering applications; the name is short for FORmula TRANslation; Fortran is still in use today by scientists because of its very capability to carry out numeric computation quite efficiently.
Genstat	a comprehensive statistics system which offers ease-of-use for the novice user through a Windows menu interface, or power and flexibility for the more experienced user through a powerful command language interface. Genstat was originally conceived and developed at the Rothamsted Experimental Station (RRES, UK), approximately 30 years ago.
IPSC	Institute for the Protection and Security of the Citizen, providing

	research-based, systems-oriented support to EU policies so as to protect the citizen against economic and technological risk. The Institute maintains and develops its expertise and networks in information, communication, space and engineering technologies in support of its mission. The strong cross-fertilisation between its nuclear and non-nuclear activities strengthens the expertise it can bring to the benefit of customers in both domains.
IMSL	International Mathematical and Statistical Libraries; they are a comprehensive set of mathematical and statistical functions that programmers can embed into their software applications. The IMSL Libraries provide high-performance computing software and expertise needed to develop and execute sophisticated numerical analysis applications. These libraries free users from developing their own internal code by providing pre-written mathematical and statistical algorithms that can be embedded into computer applications.
JRC	Joint Research Centre: a research based policy support organisation and an integral part of the European Commission, providing independent scientific and technical advice to the Commission and EU Member States in support of European Union (EU) policies. Main aim is to help to create a safer, cleaner, healthier and more competitive Europe.
MARS	Monitoring Agriculture with Remote Sensing, a project started in 1988, initially designed to apply emerging space technologies for providing independent and timely information on crop areas and yields. Since 1993, driven by user requirements, the team has contributed towards a more effective and efficient management of the Common Agricultural Policy through the provision of a broader range of technical support services to DG Agriculture and Member State Administrations.
MARS-STAT	The objective of MARS STAT is to provide independent, homogenous quantitative crop statistics at EU and National levels, in near real time and within an operational system. To reach the objective, MARS STAT has developed methodologies for early crop yield and area estimates, such as the MARS Crop Yield Forecasting System, based on satellite information, agro-meteorological modeling and statistical analyses
MARSOP	MARS Operational: the project which was carried out by the MARS consortium led by Alterra in the period 2000-2003 and which is now continued in the next term 2004-2008, in order to provide early information on the development and growth conditions of crops.
Mallows Cp	A measure of goodness-of-prediction. In general, one should look for models where Mallows Cp is small. A small Cp value indicates that the model is relatively precise (has small variance) in estimating the true regression coefficients and predicting future responses.
MCYFS	MARS Crop Yield Forecasting System; was instated as part of the MARS activities to supply the DG Agriculture and EUROSTAT

	with early information on development, growth conditions and expected yields of crops
NOAA/AVHRR	A type of sensor on board of the NOAA satellites, called Advanced Very High Resolution Radiometer; these satellites were operated by a service of the National Oceanic and Atmospheric Administration of the USA.
NUTS	Nomenclature of Units for Territorial Statistics; a coding system used especially by EUROSTAT for reporting statistics on all the administrative regions of the EU.
ODBC	Open DataBase Connectivity, a standard database access method developed by the SQL Access group in 1992. The goal of ODBC is to make it possible to access any data from any application, regardless of which database management system (DBMS) is handling the data. ODBC manages this by inserting a middle layer, called a database driver , between an application and the DBMS. The purpose of this layer is to translate the application's data queries into commands that the DBMS understands.
R-squared	A mathematical term describing how much variation is being explained by the X's in a regression model. The R-squared value is the fraction of the variance in the data that is explained by a regression model.
RS	Remote Sensing: in the broadest sense, this is the measurement or acquisition of information of an object or phenomenon, by a recording device that is not in physical or intimate contact with the object. In practice, remote sensing is the utilization at a distance (as from aircraft, spacecraft, satellite, or ship) of any device for gathering information about the environment. In modern usage, the term usually refers to techniques involving the use of instruments aboard aircraft and spacecraft.
S-Plus	Platform for statistical analysis. The basis of this platform is the S programming language which was specifically developed for the creation of analytic prototypes. It is an interactive language, allowing statisticians and developers to compare multiple models and share results across systems.
VIF	Variance Inflation Factor. It measures the impact of collinearity among the X's in a regression model on the precision of estimation. It expresses the degree to which collinearity among the predictors degrades the precision of an estimate. Typically a VIF value greater than 10 is of concern.
WOFOST	WOFOST is a mechanistic model that explains crop growth on the basis of the underlying processes, such as photosynthesis and respiration, and how these processes are affected by environmental conditions. The model describes crop growth as biomass accumulation in combination with phenological development. It simulates the crop life cycle from sowing or emergence to maturity. Meteorological data (rain, temperature, wind speed, global radiation, air humidity) are needed as input. Other input data include

	volumetric soil moisture content at various suction levels, and other data on saturated and unsaturated water flow. Also data on site specific soil and crop management are requested.