Crop Monitoring as an E-agricultural tool in Developing Countries

CGMS STATISTICAL TOOLBOX

Reference: *E-AGRI_D61.1_CGMS_Statistical_Toolbox_F1.0*

Author(s): Allard de Wit, Steven Hoek

Version: 1.0

Date: 7/04/2014

DOCUMENT CONTROL

Signatures

Author(s) : Allard de Wit, Steven Hoek

Reviewer(s) : Qinghan Dong

Approver(s) :

Issuing authority :

Change record

Release	Date	Pages	Description	Editor(s)/Reviewer(s)
1.0	07/04/2014	27	Report on CST implementation for the test sites.	Allard de Wit Qinghan Dong

TABLE OF CONTENT

LIST OF FIGURES

ACRONYMS & GLOSSARY

AIFER	-	Anhui Institute For Economic Research
CGMS	-	Crop Growth Monitoring System
CST	-	CGMS Statistical Toolbox
DMN	-	Direction de la Météorologie Nationale
DSS	-	Direction de la Statistique
INRA	-	Institut national de la recherche agronomique
PCA	-	Principle Component Analysis
WOFOST	-	WOrld FOod STudies

EXECUTIVE SUMMARY

The work package 6.1 focused on the implementation of the CGMS Statistical Toolbox (CST) for the two target regions (Morocco and Anhui, China). The CST is a tool that can analyse the relationship between historical regional reported crop yields and a set of indicators. If a sufficiently accurate relationship is found, it can then be used for crop yield forecasting purposes. We can conclude that The CGMS Statistical Toolbox has been implemented successfully for the two target regions. Besides the implementation of the database and the real-time flow of indicators a large number of improvements have been made to the application itself.

For the Morocco test site, the CST has been embedded into the processing chain at DMN and implemented in an ORACLE Database. Simulation results from the WOFOST crop simulations and aggregated meteorological variables are inserted as indicators into the CST database at the end of each dekad. Moreover, satellite-based indicators that are available from INRA were included into the processing chain and were added to the CST database as well. This means that the CST users at DMN, INRA and DSS can use the full spectrum of indicators (meteo, crop simulations, satellite) for their analyses and for making the final yield forecast. The users at the above-mentioned institutes have received a thorough training on the background and use of the CST as part of WP6.2.

For the Anhui test site, the CGMS processing chain was simpler and based on a Microsoft Access database. The CST database was therefore also based on Microsoft Access and implemented as part of the entire processing. In Anhui, the aggregated WOFOST simulation results were inserted into the CST database at the end of each dekad through the CGMS executable. Training in the use of the CST was done as part of a visit of AIFER personnel to Alterra.

The overall conclusion is that the CST has demonstrated to be a valuable tool in analysing historical regional crop yields and combining them with crop indicators for making crop yield forecasts. CST is easily explained to end users and provides clear information to its users on the performance of the available indicators for yield forecasting. Finally, the CST formalizes the work flow for making yield forecasts and the use of CST is much less prone error as manual analysis through general statistical software packages (Excel, R, SAS, etc.).

1. Introduction

1.1. Introduction to regional crop yield forecasting

A general characteristic of systems for regional crop monitoring is that they generally do not predict the crop yield directly. Instead, these systems predict indicator values that are known (or assumed) to be correlated with the regional reported crop yields. The reason is that the true farmer's yield at the field level or a regional aggregated value cannot be accurately predicted using satellite data, models or meteorological variables. There can be either large biases in the indicator values (simulation models) or the values that are provided by the monitoring system are not even in the same unit of measurements, e.g. precipitation sum in mm or dimensionless satellite vegetation indices vs crop yields in ton/hectare. Moreover, the reported regional crop yields often contain trends due to technological advancement that have to be taken care of when forecasting the crop yield.

To illustrate this situation, Figure 1 shows the CGMS water-limited crop yield at the end of the growing season and the EUROSTAT reported yield for wheat in Spain and sugar-beet in Germany. From these figures a number of conclusions can be drawn. First of all, the reported yields for the two examples contain a trend of rising crop yields in time. For wheat in Spain yields have increased from 1.64 ton/ha (1975—1979 average) to 2.83 ton/ha (2005-2009 average). For sugar beet in Germany yields have increased from 46.6 ton/ha (1988—1992 average) to 61.8 ton/ha (2005-2009 average). In both countries the trend over the entire window is significant at $\alpha = 0.001$.

This trend is generally called the 'technology trend' and is caused by improved agricultural practices (with regard to yield) over time. The CGMS simulated results do not contain such trend as the technology level is assumed stable. Although in many European countries the direction of the trend is positive, there are examples of negative trends in many countries for example due to decline in soil fertility or where the cultivation of commodity crops is being pushed onto marginal lands.

Secondly, there is a considerable mismatch in absolute terms between CGMS simulated yields and EUROSTAT reported yields. For wheat in Spain average simulated yield is 5.1 ton/ha and average reported yield is 2.3 ton/ha. For sugar-beet in Germany average simulated yield is 13.0 ton/ha and average reported yield is 54.4 ton/ha. For wheat in Spain, the difference is caused by sub-optimal management by farmers, while CGMS assumes that management is optimal in terms of nutrients, pests and diseases. For sugar-beet in Germany the difference is mainly caused by the water content of the reported yield (fresh weight) while WOFOST predicts dry weight.

Finally, the inter-annual variability matches quit well between the time-series of reported and simulated yields. Particularly the extreme years can be recognized in both time-series easily: in Spain the 1995 and 2005 drought years, in Germany the 2003 and 2006 drought years and the favourable years 1993 and 2000. Nevertheless, particularly for wheat in Spain, the CGMS simulated yields show much larger fluctuations in yield than the reported yields.

Based on this example, we can conclude that we need a "mapping" to convert output from regional crop monitoring systems to a forecast of the actual crop yield. This is exactly the objective of the CGMS Statistical Toolbox.

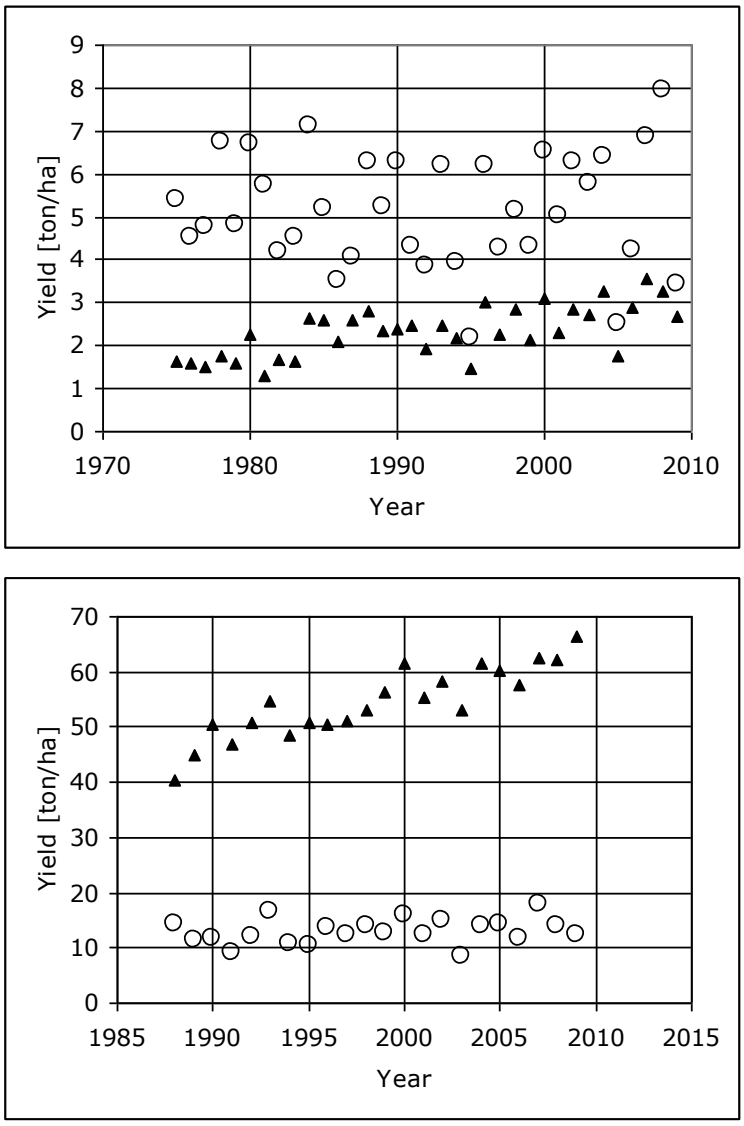


Figure 1. Time-series of CGMS simulated yields (○) and EUROSTAT reported yields (▲) for wheat in Spain (upper) and sugar-beet in Germany (lower).

1.2. Role of the CGMS statistical Toolbox

The CGMS Statistical Toolbox (CST) is designed to build the "mapping" between the regional crop yield statistics and an arbitrary set of indicators that is assumed to have some correlation with the reported yields. Often these indicators are derived from crop monitoring system and consist of crop model simulation results, satellite observations or meteorological values. However, other indicators that are deemed valuable (and can be obtained before the harvest) can be added. For example, the amount of fertilizer ordered by farmers could be an important indicator and the CST would be able to handle such an indicator as well.

The overall approach used by the CST is that it searched for a relationship between the historic reported regional crop yield and the historic set of indicators that is available in the database. As the validity of the relationship must be tested, it is important that the length of the historic records is sufficiently large. In practice this means that 10 to 15 years of historical data should be available in the database to build a mapping with sufficient confidence. When such a mapping is found, it is assumed that his mapping is also valid for the current crop growth season and it can be used to predict the current year crop yield from the indicator values provided by the monitoring system.

1.3. Approaches implemented in the CST

The CST contains three approaches that can be used for building the mapping between historic regional reported crop yields and the indicator values provided by the monitoring system. First of all, the CST provides an interface for trend analysis that can be used to identify whether the time-series of reported regional yield contains a trend, and if so, what the best model is to fit the trend (either linear or quadratic). The CST also supports the selection of the length of the period for which the trend model is valid. Visualisation of this trend with the statistical data also helps to detect any trend breaches that can occur in the time-series. For example, time-series of European regional crop yield contain trend breaches due to changes in the Common Agricultural Policy of the European Commission.

Second, the CST provides (multiple) linear regression to correlate one or more indicator variables with the historical reported yields including the trend model that was selected (no trend, linear or quadratic). Besides carrying out the regression itself, the CST provides important information about the significance of the different indicators in the regression model. For example, it is well known that regression models with more indicators have a tendency to inflate the correlation of the results. This means that the model with more indicators may fit better on the historical results, but does not provide an improved capacity for prediction. The CST handles this effect and provides graphical and quantitative information to the end user about the significance of each term in the regression model.

Finally, the CST provides a module for scenario analysis. In this analysis, the focus is not on correlation but on similarity between the current year and the historic record. The drawback of a regression analysis is that it assumes linear correlation between indicators and the historic reported yields which may not exist in practice (for example due to extreme events). Similarity analysis focuses on finding those years in the historical records which are similar to the current one

given the indicator values available. For this purpose, a principle component analysis (PCA) is carried out to reduce the dimensionality of the input indicator values. A selected number of principal components are then used to select the years which are in the neighbourhood (in PCA space) of the current year. Those years are than selected and the (weighted) average of the crop yield in the selected years is used as a prediction for the current year, possibly adjusted for a trend by the trend model.

The statistical approaches implemented in the CST could be carried out with any package for statistical analysis (R, SAS, etc.). Nevertheless, it has been shown over the years that this is prone to error. The advantage of CST is that it provides a dedicated interface and work flow for making such analysis and provides quality control on the regression models that are found. Finally, user-defined settings for the selected forecasting model can be stored in CSV files and can be used as input into the *CST batch mode* which automatically produces the yield forecast based on the settings read from the CSV file.

1.4. Application of CST within E-Agri

As outlined above, the CGMS Statistical Toolbox provides the link between output from crop monitoring systems and the final crop yield forecast in physical quantities at regional level. Within the framework of the e-Agri project, it is logical that also the CST becomes an integral part of the processing chains that have been developed for the test regions.

During the project, the CST has been implemented for the processing developed within Work Package 2 for the test regions in Morocco and Anhui (China). Moreover, the CST has been implemented in the BioMa application for Jiangsu (rice) and Morocco (wheat) in WP3. The implementation of the CST for these test regions in WP2 covered the following actions:

- Modification of the CST database to cover the test regions;
- Preparing and loading of the regional statistics for the target regions;
- Implementation of the CST within the processing chain in order to have real-time updates of the indicators that CST uses to perform the historical analysis and the crop yield forecast.

1.5. General improvements to CST

Within the framework of e-Agri a number of general improvements and bug-fixes were carried out on the CGMS Statistical Toolbox that are not specifically related to any test region (see Annex 1). Moreover, the database structure that is used by CST was adapted to the new CGMS11 schema.

2. Implementation of CST for Morocco test site

2.1. Filling of the CST database for Moroccan test site

2.1.1. Regions and hierarchy

The CST needs to know the regions that are located in the test site and their hierarchy (e.g. which districts belong to which province). For the Moroccan test site the regions at the lowest level are the districts, above that are provinces, Agro-Zones and finally Morocco as a nation. The entire Moroccan territory is covered by 1490 districts, 45 provinces and 6 Agro-Zones.

The consistency of the region hierarchy was guaranteed by starting with the district boundaries and adding attribute values to each district that referred to the ID of the province and Agro-Zones where this district was located. After the assignment of attributes, the maps of province, Agro-Zones and Morocco were derived by applying a GIS procedure that dissolves boundaries between districts if they have common province ID, between provinces if they have a common Agro-Zone ID and finally all Agro-Zones into one map for Morocco.

By visualizing the resulting maps in a GIS, errors in the assignment of IDs for a district could be easily recognized and corrected. This procedure was repeated several times, until the map at all levels was correct. Finally, the database table in the CST database was loaded with all regional entities, its level in the hierarchy and the region to which it belongs (In total 1491 records).

2.1.2. Historical reported crop yields

The historical regional reported crop yields were provided by DSS and consisted of crop production and cultivated area for each province in Morocco. Based on this data, the crop yield for each province was calculated (production divided by area) at province level and aggregated up to Agro-Zones and national level. Finally, all historical data was loaded into the relevant database table for CST. See Table 1 for an overview of the statistical data for Morocco

Table 1. Overview of statistical data available for Morocco

Crop	Nr provinces	Start year	End year
Soft-wheat	41	1979	2012
Durum-wheat	41	1979	2012
Barley	42	1979	2012

2.2. Implementation of the data flow in the CGMS processing chain

The implementation of the processing chain for the CST was already described in the deliverable D25.1 CGMS Piloting report and will be repeated here for completeness

The aggregated CGMS crop simulation results at regional level are one the of indicator sets which are used for crop yield forecasting by the CGMS statistical toolbox. At DMN a second database has been created which holds the database schema necessary for storing the data for the CGMS Statistical and results from the application. The entire flow of information is depicted in Figure 2.

Results from the CGMS crop simulation at regional level are send to the CST database schema which can accomplished through several select-insert statements as this is the same database. At the same time, additional indicators are added which are derived from the weather data directly. These include temperature sum and precipitation sum which are important indicators in Morocco as well. As the CGMS executable does not calculate weather indicators at regional level, calculation of weather indicators at regional level is accomplished by a separate ORACLE package (the so-called CMETEO package). Finally, the satellite-based indicators are provided by the satellite processing chains from partner VITO and are available at INRA. Therefore, these indicators are provided by INRA through ftp and are integrated into the CST database at DMN.

The entire CST database in ORACLE is then replicated in a Microsoft Access database which can be easily done due to the strong data reduction that is accomplished by the aggregation from grid level to the regional level. Finally, this Microsoft access database is packaged (zipped) and put on a medium for file-sharing such as dropbox or an ftp server that can be accessed by CST users. The Access database can be picked from the file sharing medium, unzipped onto the right CST folder structure and the end user can analyse the latest indicators for yield forecasting.

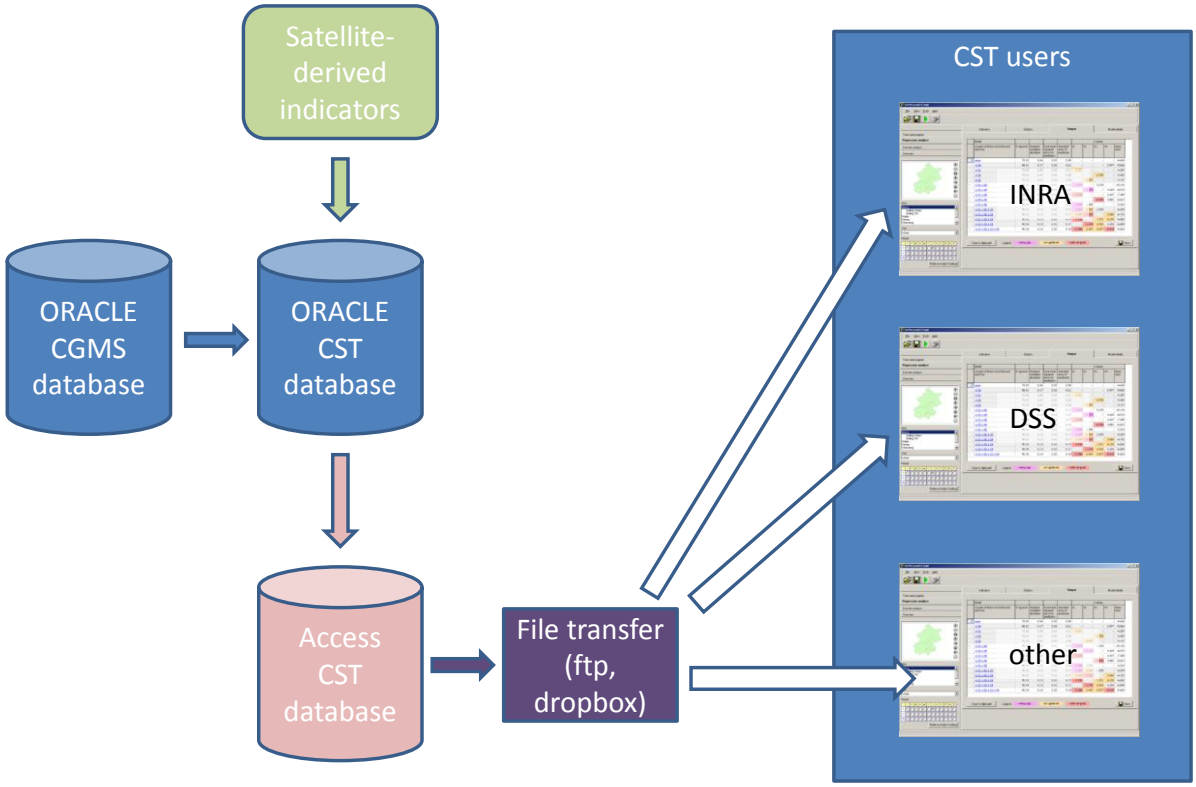


Figure 2. Flow of information from CGMS to the CGMS statistical toolbox and its end users.

2.3. The CGMS Statistical Toolbox for Morocco

Finally, the CST executable was provided as a MS Windows installer package which could be installed at the user locations at the premises of DMN, INRA and DSS and other users in the future. Figure 3 to Figure 5 show some screen shots of the CST setup for Morocco.

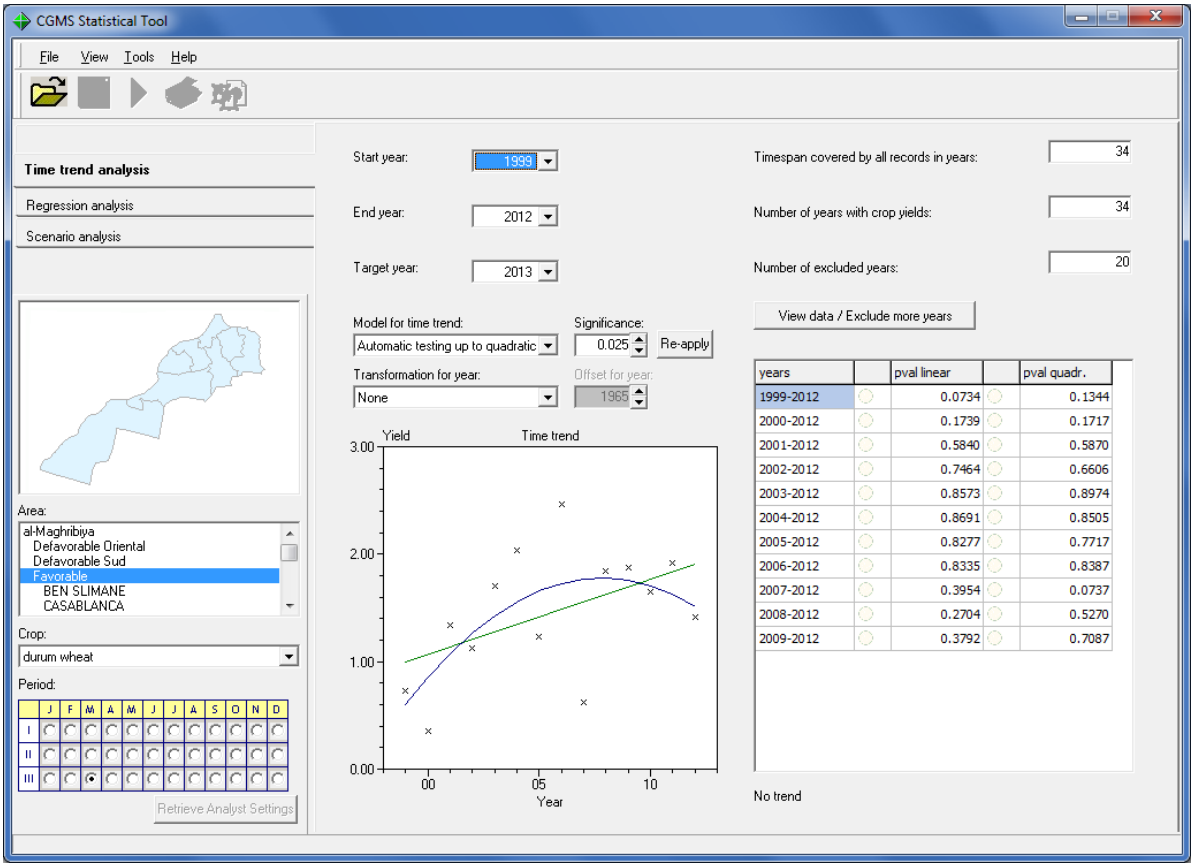


Figure 3. The Time-trend analysis window for the CST Morocco implementation.

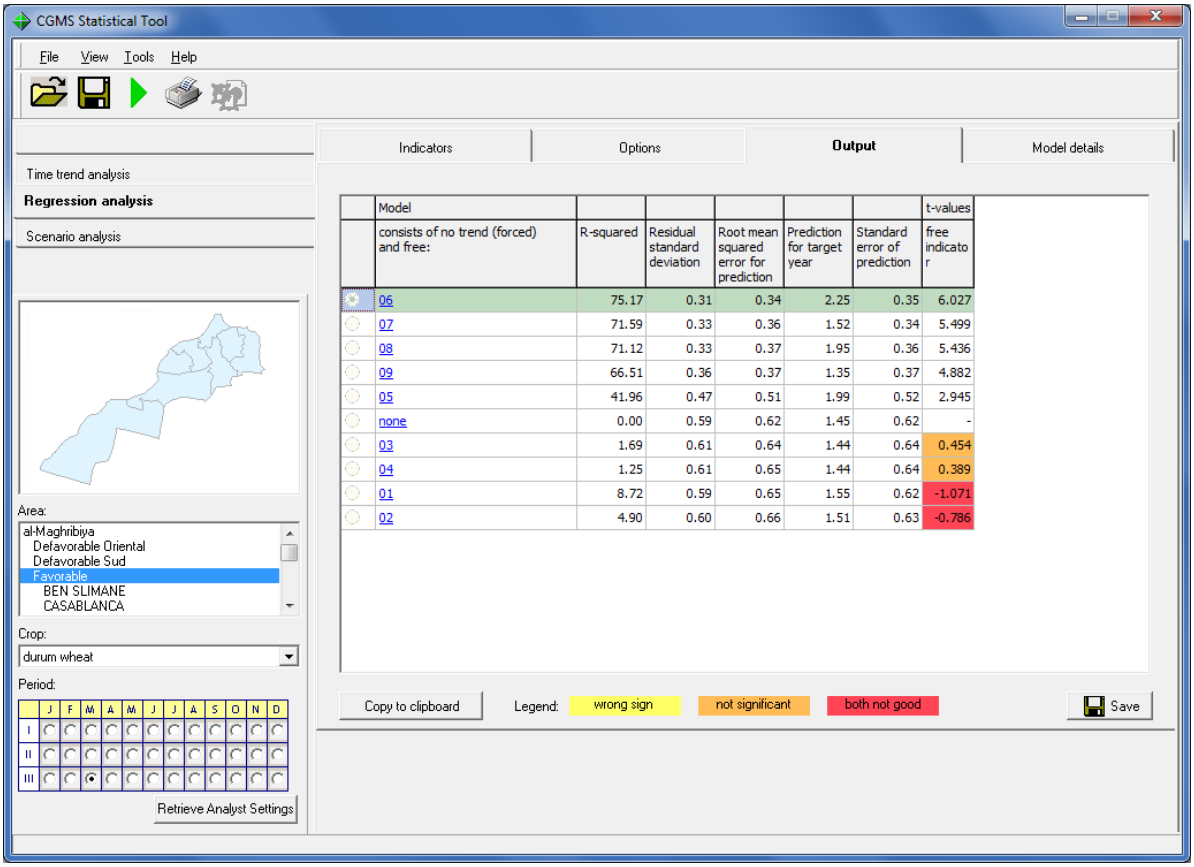


Figure 4. The regression analysis results window for the CST Morocco implementation.

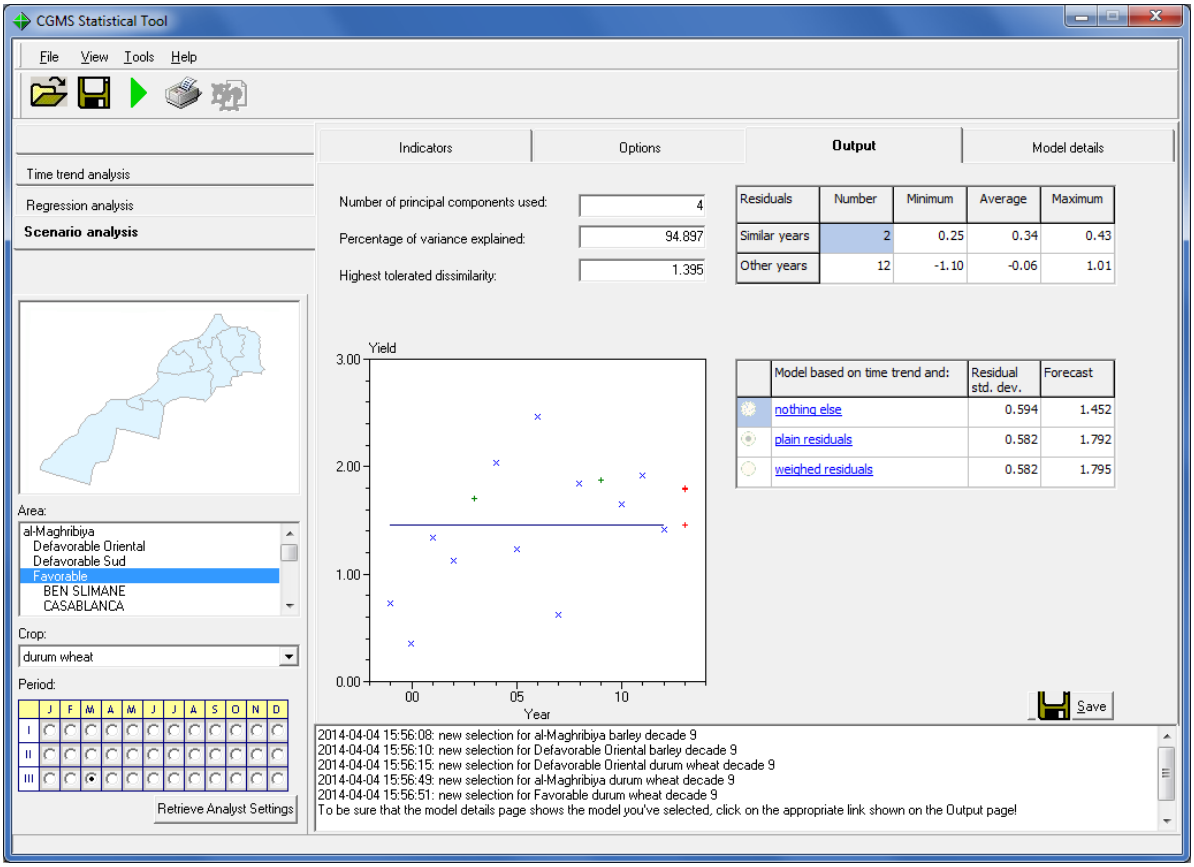


Figure 5. The scenario analysis results window for the CST Morocco implementation.

3. Implementation of the CST for Anhui, China.

3.1. Filling of the CST database for the Anhui test site

3.1.1. Regions and hierarchy

The regional division in Anhui was processed in exactly the same way as in Morocco which lead to a consistent hierarchical division albeit on three levels (county, district and province) instead of four. For the Huaibei plain Anhui, the regional division consisted of 23 counties, 6 districts up to the level of whole Anhui (in practice covering the Huaibei Plain region because this is the winter-wheat cultivation area).

3.1.2. Historical reported crop yields

Historical reported production and cultivated area of winter-wheat were provided by AIFER at the level of districts. These statistical data were used to calculate the crop yield at district level and at the level of Anhui. The time-series covered the period 2000 to 2011.

3.2. Implementation of the data flow in the CGMS processing chain

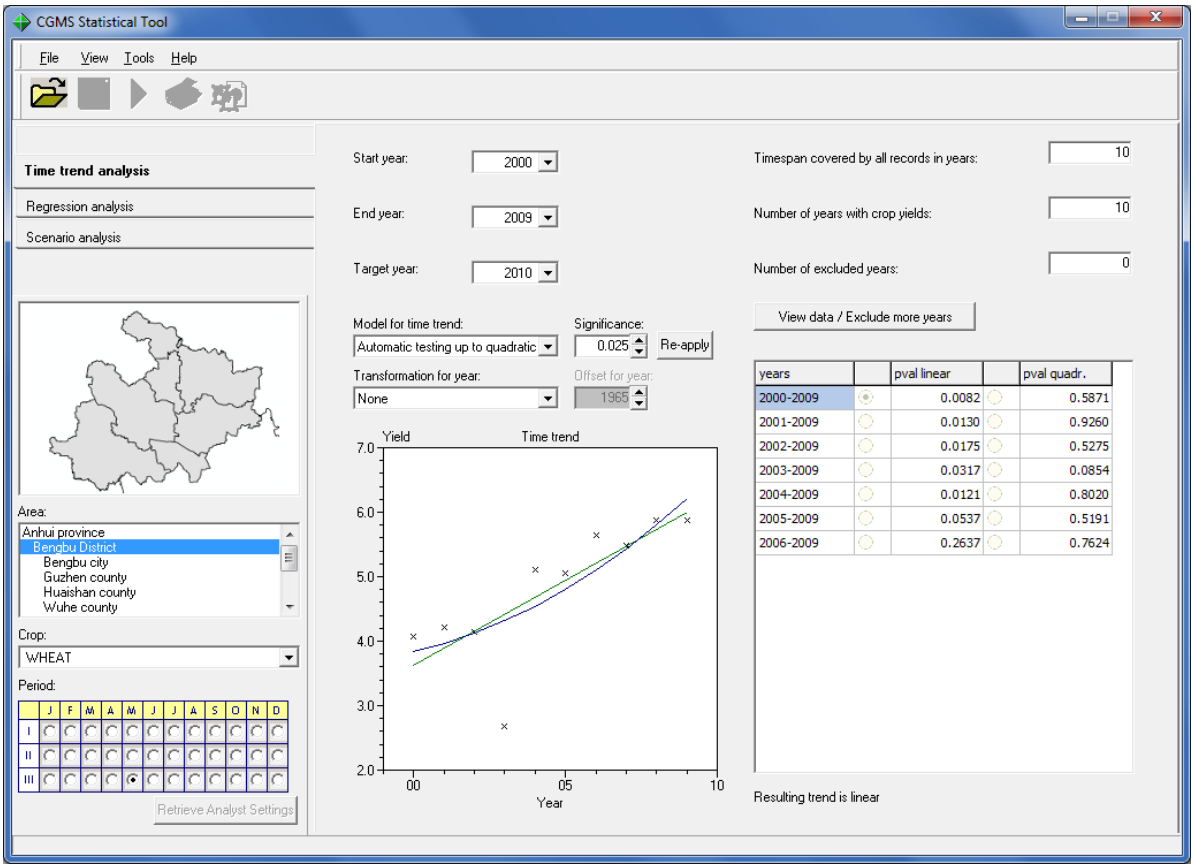
The processing steps that were implemented in the CGMS Anhui processing chain have already been described in the deliverable D24.1 “CGMS Anhui piloting report” and will be reported here for completeness.

All processing in the CGMS-Anhui is carried through the CGMS executable because the underlying database is a Microsoft Access database. The procedures for the CGMS Statistical Toolbox present the final step in the CGMS processing chain: the aggregation of results to the administrative regions and the preparation of results at administrative level for the yield forecasting procedures. These are implemented in two procedures in the Windows Command Language which execute the CGMS executable:

1. Aggregation of the crop simulation results to administrative level can be started from ‘1_runcgms_aggregation.bat’. This will read the configuration from the file ‘cgms_aggregation.ini’. The settings in the .INI file need to be set manually, but this can be automated easily in the future.
2. Preparation of the results and inserting them into the CST database for analysis with the CGMS Statistical Toolbox (WP6) can be started with the script ‘2_runcgms_preparation.bat’ which reads the configuration from the file ‘cgms_preparation.ini’. The settings in the .INI file need to be set manually, but this can be automated easily in the future.

3.3. The CGMS Statistical Toolbox for Anhui

Finally, the CST executable was provided as a MS Windows installer script which could be installed at the user locations at the premises of AIFER and other users in the future. Figure 6 to Figure 8 show some screen shots of the CST setup for Anhui.



years	pval linear	pval quadr.
2000-2009	0.0082	0.5871
2001-2009	0.0130	0.9260
2002-2009	0.0175	0.5275
2003-2009	0.0317	0.0854
2004-2009	0.0121	0.8020
2005-2009	0.0537	0.5191
2006-2009	0.2637	0.7624

Figure 6. The Time-trend analysis window for the CST Anhui implementation.

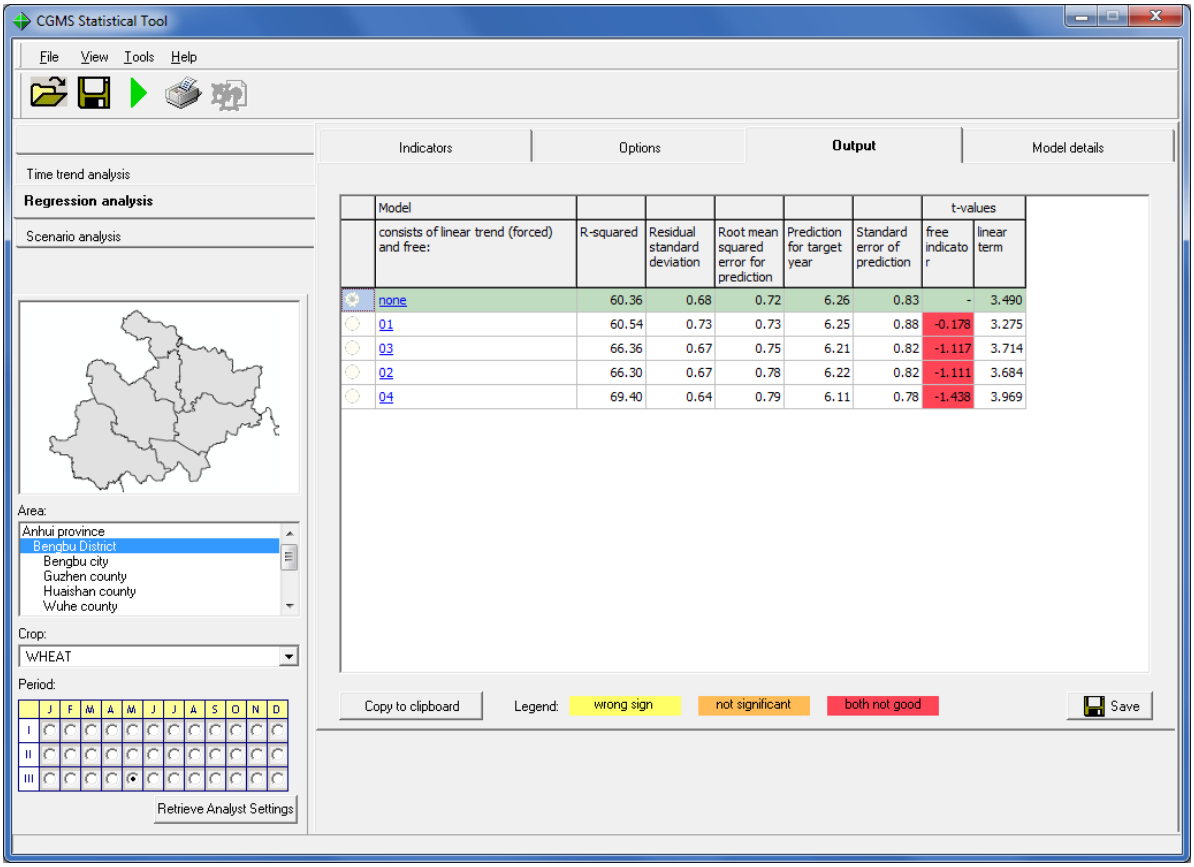


Figure 7. The regression analysis results window for the CST Anhui implementation.

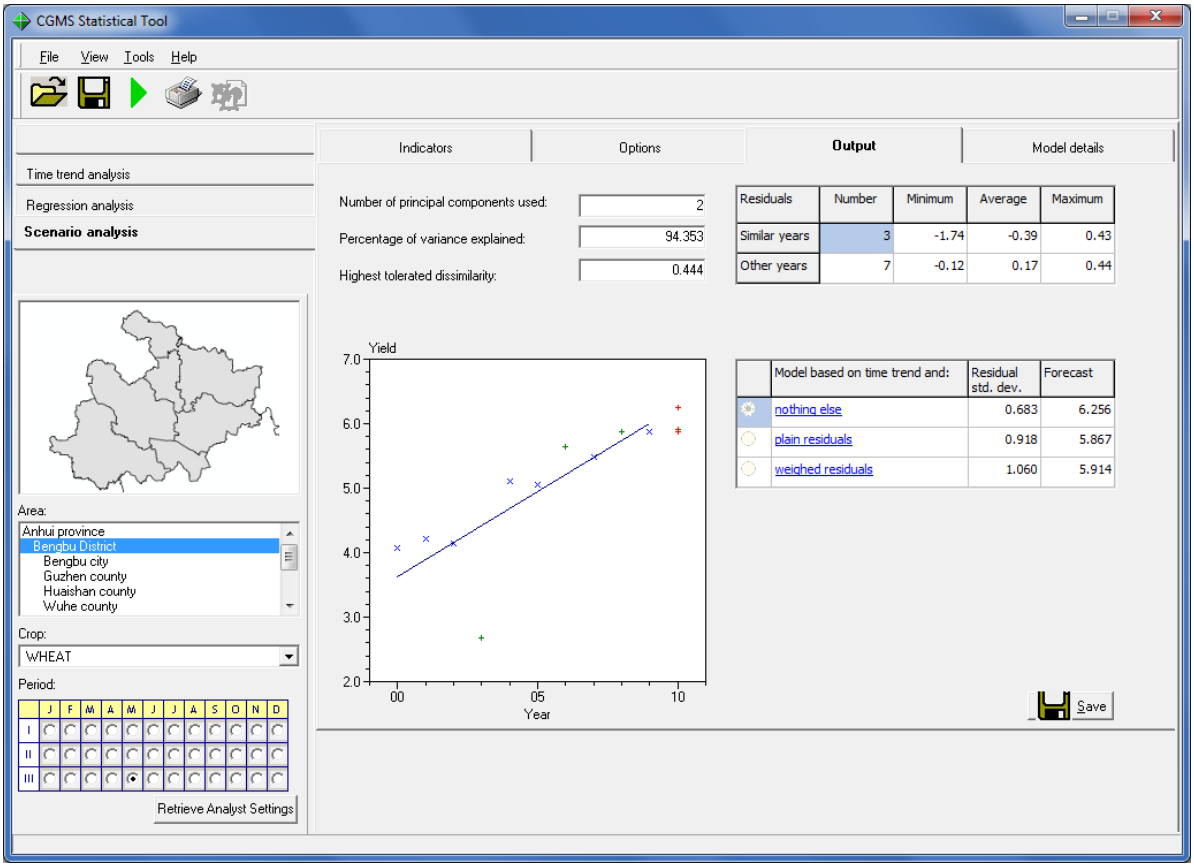


Figure 8. The scenario analysis results window for the CST Anhui implementation.

4. Conclusions

The CGMS Statistical Toolbox has been implemented successfully for the two target regions. Besides the implementation of the database and the real-time flow of indicators a large number of improvements have been made to the application itself.

For the Morocco test site, the CST has been embedded into the processing chain at DMN and implemented in an ORACLE Database. Simulation results from the WOFOST crop simulations and aggregated meteorological variables are inserted as indicators into the CST database at the end of each dekad. Moreover, satellite-based indicators that are available from INRA were included into the processing chain and were added to the CST database as well. This means that the CST users at DMN, INRA and DSS can use the full spectrum of indicators (meteo, crop simulations, satellite) for their analyses and for making the final yield forecast. The users at the above-mentioned institutes have received a thorough training on the background and use of the CST as part of WP6.2.

For the Anhui test site, the CGMS processing chain was simpler and based on a Microsoft Access database. The CST database was therefore also based on Microsoft Access and implemented as part of the entire processing. In Anhui, the aggregated WOFOST simulation results were inserted into the CST database at the end of each dekad through the CGMS executable. Training in the use of the CST was done as part of a visit of AIFER personell to Alterra.

The overall conclusion is that the CST has demonstrated to be a valuable tool in analysing historical regional crop yields and combining them with crop indicators for making crop yield forecasts. CST is easily explained to end users and provides clear information to its users on the performance of the available indicators for yield forecasting. Finally, the CST formalizes the work flow for making yield forecasts and the use of CST is much less prone error as manual analysis through general statistical software packages (Excel, R, SAS, etc).

Annex 1: general improvements and bug fixes to the CST

The following application improvements and bug fixes were carried out on the CGMS Statistical Toolbox:

- Represent more accurately how many of the total years in the selected periods have valid crop yield data
- Solve the bugs that were found in the new screen “Saved Models”
- Add a “Save” button as well as an “Export” button to the Output screens (for regression and scenario analysis)
- Solve two memory problems
- Prevent double log messages
- Replace the code for representing the geographical hierarchy with new code based on a better concept – after problems were established
- To find a workaround for a representation problem that occurs in certain cases when a CSV-file with settings is opened
- Make sure that the batch mode does or does not write the predicted crop yield to the database, depending on a value in the ini-file
- Make the batch mode less sensitive to deviant input values in the CSV-files
- To establish which setting is used in the batch mode: OrderOfTimeTrend or TrendModelType
- Solve a problem with the models that have a logarithmic transformation of the years built in
- Links represented in the output screens now contain an additional tag which is supposed to confuse Internet Explorer so that the generated HTML documents are not stored in the cache; however, in some cases those additional tags are not needed and rather had to be filtered out
- Add values for the *F*-statistic and its probability to the model details documents
- Make sure that the list with summary statistics on the Options screen for regression analysis does not get filled with double items – not even in the case a CSV-file with settings is opened and a regression model is loaded
- An auxiliary unit was split into three in order to separate things better conceptually: GeneralModel, RegressionModel and ScenarioModel.

- Modify the batch mode so that the results of standalone and batch modes are comparable
- Make it possible to use Corrected Mallows C_p also as selection criterion in batch mode
- Prevent “Out of bounds” error in the unit that retrieves the indicator data from the database
- A few paragraphs in the documentation were revised.