

The CGMS crop yield forecasting system

Steven Hoek & Allard de Wit



Introduction, part 1

WUR = Wageningen University & Research Centre

Legal entity behind the research centre: Stichting DLO

DLO is divided up into more than 5 institutes, including:

- Alterra (environmental science)
- Plant Research International (plant science), includes Biometris

Introduction part 2

Alterra:

- Centre for Geo-Information
- Centre for Water and Climate
- Centre for Soil
- Centre for Landscape
- Centre for Ecosystems

Contents

- Some background on crop yield forecasting
- Yield forecasting models
- The CGMS Statistical Toolbox (Crop Yield forecasting tool) demonstration
- Play around with the CGMS Statistical Toolbox

Goals

- Get to know yield forecasting concepts
- Become familiar with the CGMS yield forecasting tool
- Be able to carry out crop yield forecasts based on results from CGMS
- Be able to add your own indicators to the CGMS toolbox

About crop yield forecasting

- Estimates of crop yield or production for the current season before the harvest
- For administrative regions
- Often calibrated against past regional statistical data
- Continuous “assimilation” of data as the growing season progresses
- Improve accuracy during the growing season
- Better than a baseline forecast (i.e. average or trend)

About crop yield forecasting

- “the art of identifying the factors that determine the spatial and inter-annual variability of crop yields” (René Gommès, FAO 2003).

About those factors

“ ... **AND IF IT GETS ENOUGH RAIN, AND SUN, AND IF IT ISN'T KILLED BY HAIL, AND IF IT ISN'T DAMAGED BY FROST, AND IF WE CAN GET IT OFF BEFORE IT'S COVERED BY SNOW, AND IF WE GET IT TO THE ELEVATORS, AND IF THE TRAINS ARE RUNNING, AND IF THE GRAIN HANDLERS AREN'T IN STRIKE, AND IF ...** ”

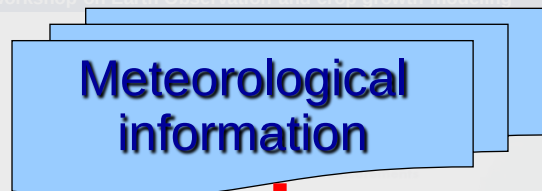


Type of forecasting systems

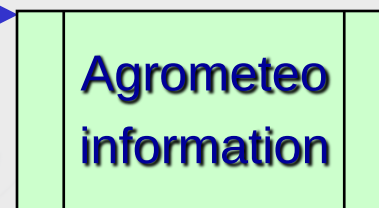
- Judgement, based on stakeholders that reach consensus on the expected yield given all available information
- Statistical, based on functional relationships between a crop yield indicator and the crop yield statistics (e.g. time trend models and/or CGMS simulation result)
- Combinations of the above (European MARS system)

MCYFS

Mixed system:
Deterministic
Statistical
Supervised



Analysts



Robust decision making

Statistical forecasting assumptions

- Uses time-series of historic statistics and crop yield indicators
- Parameterizes a forecasting model explaining the relation using a best fit criterion (mean squared error)
- The model parameters are derived at several time steps during the growing season (i.e. each dekad)
- Forecast model is then applied in prognostic mode to forecast the current season's yield

Parameterize the model in *time* or in *space*?

- ➔ Build a time-series model for one region and multiple years
- Build a spatial model for multiple regions and one year
- Combine the two above (even more difficult!)

Reasons for preferring a time-series model

Several effects:

- Socio-economic factors differ between regions
(example: Germany 7.24 ton/ha, Poland 3.44 ton/ha).
- Crop yields often show an upward (or downward) trend over the years
- Simulated year-to-year variability in crop yield differs from variability in regional statistics.

Statistical forecasting models

■ Parametric models:

- 
- Regression analyses: (multiple -) regression between crop yield statistics and crop indicators

■ Non-parametric models:

- Scenario analyses: Find similar years and use these to forecast
- Neural networks: train a neural network to recognize yield-indicator relationships

Time-series regression models for crop yield forecasting

Basic assumptions:

- crop yield = $f(\text{time-trend} + \text{indicators}(i_1, i_2, \dots))$
- Uses (multiple) linear regression

Advantages:

- Simple, understandable
- Hypothesis testing (statistical significance)
- Provides models with predictive power

Example of analysis for wheat in Morocco

1. Assessment of the data

Data view

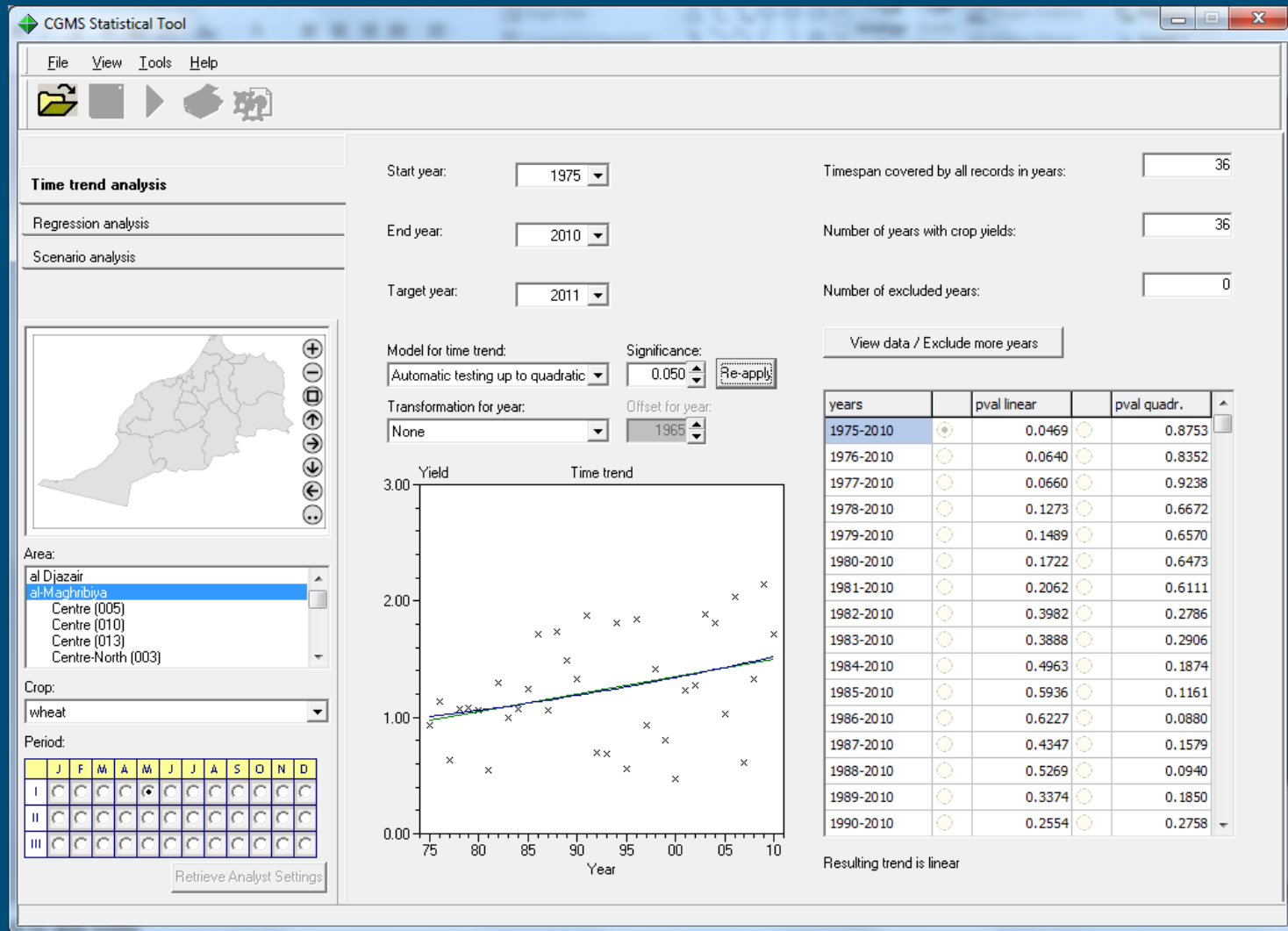
Below available data are listed for the selected area, crop and decade. You can exclude data for particular years by unchecking the boxes on the left.

Nuts area: al-Maghribiya Decade: 13
Crop: oil seed rape Included years: 1981-2006

Included	Year	Official Yield
☒	1981	1.2
☒	1982	1.2
☒	1983	1.2
☒	1984	1
☒	1985	1
☒	1986	1
☒	1987	1
☒	1988	1.364
☒	1989	0.944
☒	1990	1.021
☒	1991	0.654
☒	1992	1
☒	1993	1.1
☒	1994	0.98
☒	1995	1
☒	1996	1.533
☒	1997	1
☒	1998	1.333
☒	1999	1
☒	2000	1
☒	2001	0.75
☒	2002	1
☒	2003	1
☒	2004	1
☒	2005	1
☒	2006	1
Target	2007	*

Exclude missing  Reset  Cancel  OK

2. Time-trend analysis



3. Choose indicators

CGMS Statistical Tool

File View Tools Help

Time trend analysis
Regression analysis
Scenario analysis

Area:
al Djazair
al-Maghribiya
Centre (005)
Centre (010)
Centre (013)
Centre-North (003)

Crop:
wheat

Period:
J F M A M J J A S
I C C C C C C C C
II C C C C C C C C C
III C C C C C C C C C

Retrieve Anal

Indicators Options Output Model details

Available indicators:
Indicator name: Missing:

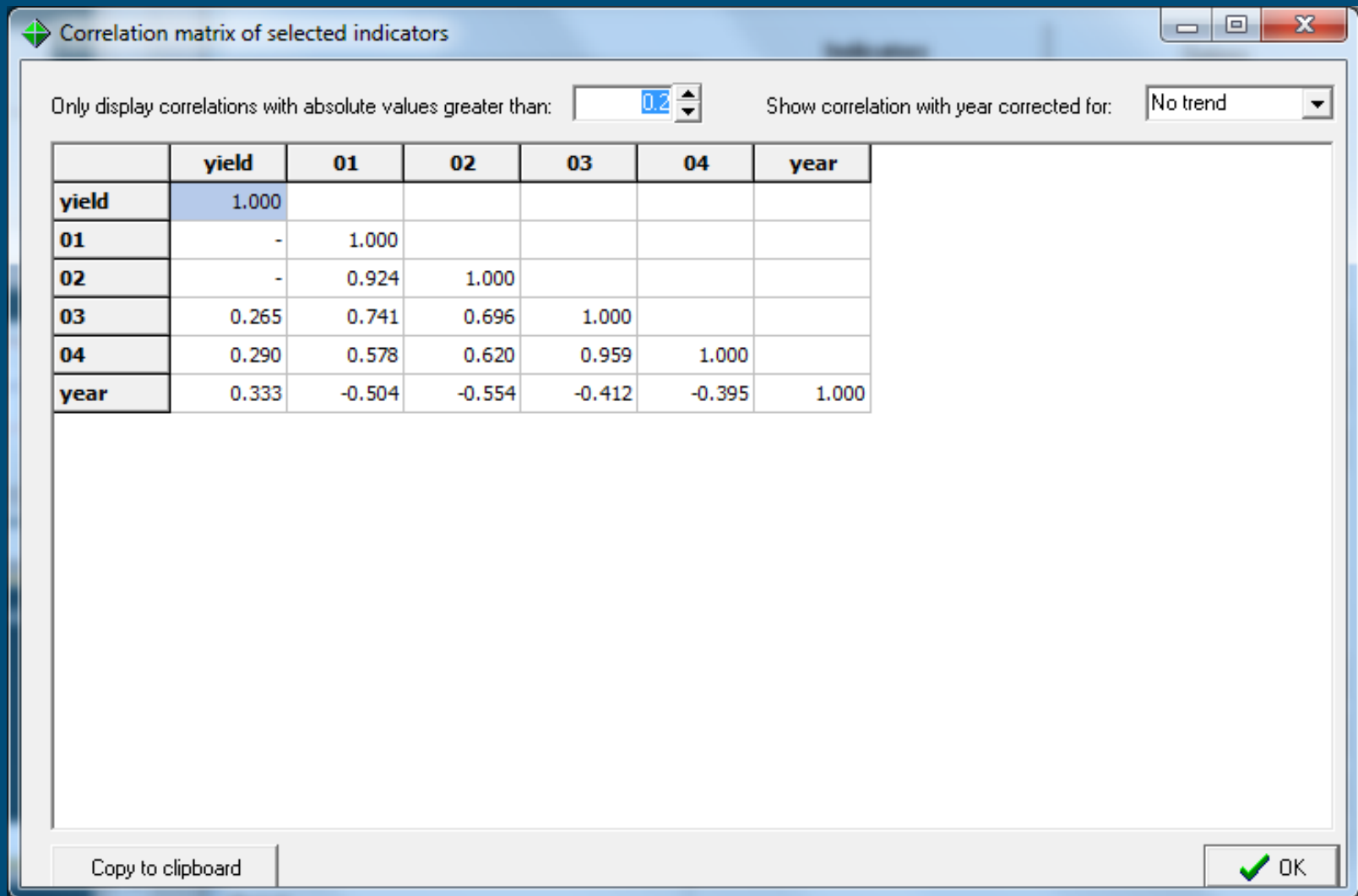
Free indicators:
Indicator name: Missing:
01 Potential Above Ground Biomass 0
02 Potential Storage Organs 0
03 Water Limited Above Ground Biomass 0
04 Water Limited Storage Organs 0

Forced indicators:
Indicator name: Missing:

Correlation matrix Next

WorkSpace
2:34 mins
4:08 mins
4:29 hours
160 k/min

4. Correlation with indicators



5. Choose options for regression analysis

The screenshot shows the CGMS Statistical Tool interface. The 'Regression analysis' tab is selected. The 'Options' section is active, showing the following settings:

- ☐ Single free indicators
- ☒ Best subset selection
- Ordering of models: Root mean squared error for predict
- ☒ Highlight indicators with incorrect sign
- ☒ Highlight terms which are not significant at: 0.050
- Summary statistics to display (max. 4):
 - ☒ R-squared
 - ☒ Adjusted R-squared
 - ☒ Mallows Cp
 - ☒ Residual standard deviation
 - ☒ Residual degrees of freedom
 - ☒ Root mean squared error for prediction
 - ☒ Prediction for target year
 - ☒ Standard error of prediction for mean
 - ☒ Standard error of prediction
 - ☒ Maximum VIF of indicators
- ☐ Only display models with VIF measure smaller than: 100.0
- Maximum number of free indicators in each model: 4
- Maximum number of best models in each subset: 5

The 'Output' section shows the 'Ordering and Sign of regression coefficients' table:

Indicator name:	Positive	Negative	Unknown	
01 Potential Above Ground	☐	☐	☒	Free
02 Potential Storage Organs	☐	☐	☒	Free
03 Water Limited Above	☐	☐	☒	Free
04 Water Limited Storage	☐	☐	☒	Free
05	☐	☐	☒	
06	☐	☐	☒	
07	☐	☐	☒	
08	☐	☐	☒	
09	☐	☐	☒	
10	☐	☐	☒	
11	☐	☐	☒	
12	☐	☐	☒	
13	☐	☐	☒	

The 'Area' dropdown is set to 'al-Djazair'. The 'Crop' dropdown is set to 'wheat'. The 'Period' dropdown is set to 'I'. The 'Retrieve Anal' button is visible.

6. Select the best model

CGMS Statistical Tool

File View Tools Help

Time trend analysis

Regression analysis

Scenario analysis

Area:

al Djazair

al-Maghribiya

Centre (005)

Centre (010)

Centre (013)

Centre-North (003)

Crop:

wheat

Period:

J F M A M J J A S O N D

I

II

III

Retrieve Analyst Settings

Indicators Options Output Model details

Model	R-squared	Residual standard deviation	Root mean squared error for prediction	Standard error of prediction	t-values					linear term
					01	02	03	04		
consists of linear trend (forced) and free:										
none	11.12	0.45	0.46	0.47	-	-	-	-	2.063	
+04	32.17	0.40	0.41	0.42	-	-	-	3.200	3.400	
+03	30.56	0.40	0.41	0.43	-	-	3.039	-	3.346	
+01	13.29	0.45	0.46	0.47	0.909	-	-	-	2.235	
+02	13.02	0.45	0.47	0.48	-	0.849	-	-	2.180	
+01 +03	34.80	0.39	0.41	0.42	-1.442	-	3.249	-	2.748	
+01 +04	32.84	0.40	0.42	0.42	-0.565	-	-	3.052	2.921	
+02 +03	33.47	0.40	0.42	0.43	-	-1.184	3.136	-	2.587	
+02 +04	33.64	0.40	0.42	0.43	-	-0.842	-	3.153	2.695	
+03 +04	32.19	0.40	0.42	0.43	-	-	0.096	0.877	3.333	
+01 +03 +04	34.93	0.40	0.43	0.44	-1.142	-	0.997	-0.249	2.469	
+01 +02 +03	34.80	0.40	0.43	0.44	-0.794	-0.004	3.197	-	2.605	
+02 +03 +04	34.23	0.40	0.43	0.44	-	-0.979	0.525	0.596	2.610	
+01 +02 +04	33.85	0.40	0.43	0.44	0.311	-0.687	-	3.103	2.647	
+01 +02 +03 +04	35.82	0.40	0.44	0.45	-0.863	0.645	0.960	-0.690	2.328	

Copy to clipboard Legend: wrong sign not significant both not good Save

7. Analyse the model details

The screenshot shows the CGMS Statistical Tool interface. The left sidebar contains a map of al-Maghribiya and a list of areas. The main panel displays the 'Model details' tab with a table of model parameters and a summary statistics section.

Area:

- al-Djazair
- al-Maghribiya**
- Centre (005)
- Centre (010)
- Centre (013)
- Centre-North (003)

Crop:

wheat

Period:

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

[Retrieve Analyst Settings](#)

Indicators | **Options** | **Output** | **Model details**

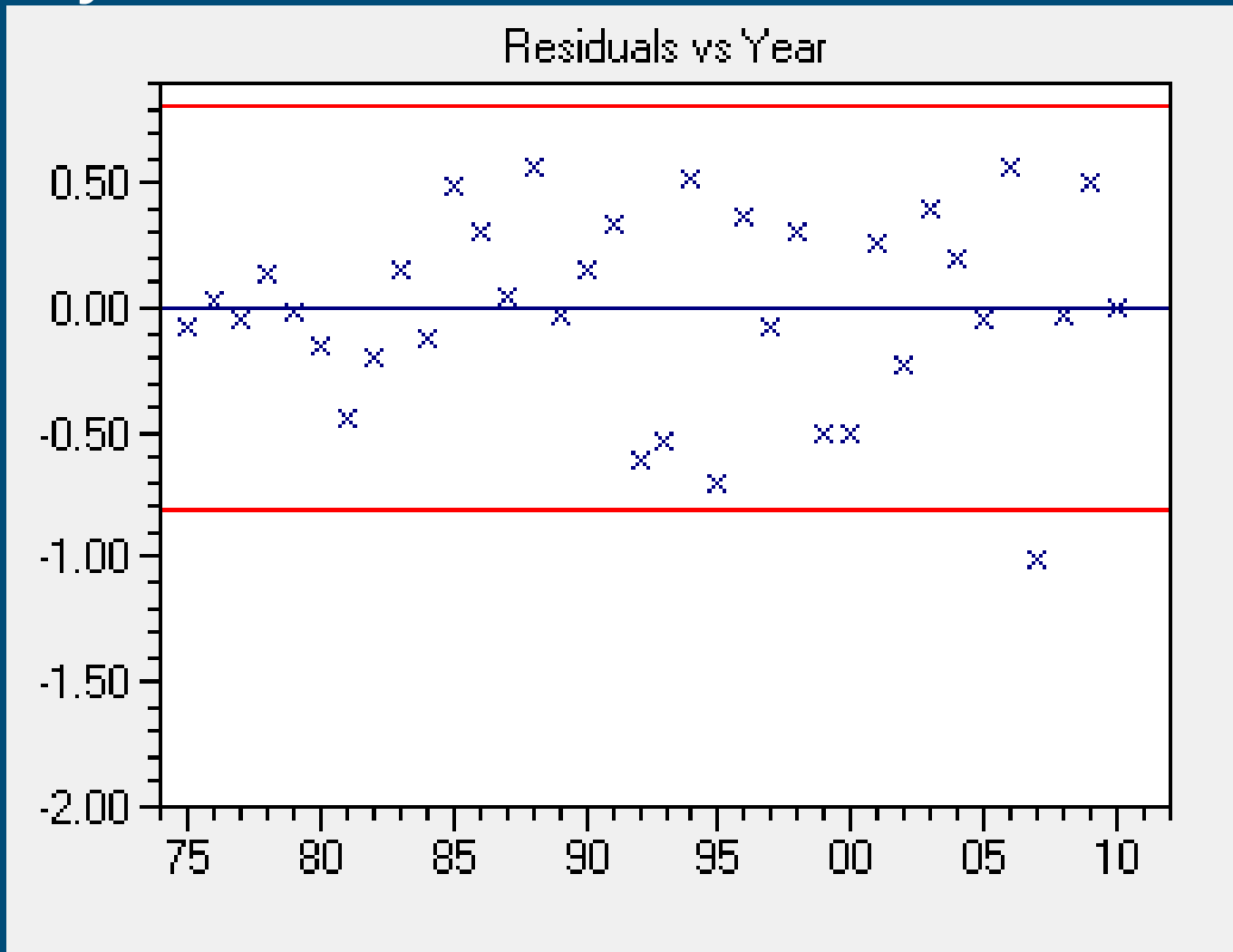
Description of model

NUTS area	al-Maghribiya
Crop	wheat
Decade	13
Included years	36
Start Year	1975
End Year	2010
Target Year	2011
Transformation of years	None
Offset for years	1965.00
Excluded years	-
Timetrend	Linear
Included indicators	04 Water Limited Storage Organs

Summary Statistics

R-squared	32.171
-----------	--------

8. Analysis of residuals



9. Correct model by excluding one/more year(s)

Data view

Below available data are listed for the selected area, crop and decade. You can exclude data for particular years by unchecking the boxes on the left.

Nuts area: al-Maghribiya Decade: 13
Crop: wheat Included years: 1975-2010

Included	Year	Official Yield	01 Potential Above Ground	02 Potential Storage Organs	03 Water Limited Above Ground	04 Water Limited Storage Organs
☒	1985	1.245	14920.531	9134.584	10806.201	5484.217
☒	1986	1.714	17026.259	10640.808	15565.651	9187.613
☒	1987	1.061	15212.344	9532.204	12320.478	6692.539
☒	1988	1.735	16502.842	10088.439	13894.148	7559.507
☒	1989	1.493	16796.405	10586.994	15599.219	9459.105
☒	1990	1.329	16246.237	10189.433	13056.819	7248.636
☒	1991	1.87	16550.871	9907.813	15943.374	9311.368
☒	1992	0.701	15900.447	9815.308	13639.695	7795.959
☒	1993	0.681	16298.023	9771.948	13561.638	7092.953
☒	1994	1.811	17058.591	10429.724	13952.8	7398.674
☒	1995	0.554	15079.549	9151.605	12862.273	7064.661
☒	1996	1.842	14398.321	8924.108	13661.818	8188.561
☒	1997	0.929	14372.021	8828.763	10571.652	5227.735
☒	1998	1.418	14118.485	8913.758	10909.033	5801.088
☒	1999	0.8	16933.929	9948.879	13718.114	6785.323
☒	2000	0.476	15657.947	9805.458	10189.12	4713.456
☒	2001	1.228	15499.512	9386.249	10419.356	4527.789
☒	2002	1.279	15082.262	9306.406	13045.378	7613.847
☒	2003	1.889	15674.579	9401.927	13396.972	7328.762
☒	2004	1.808	16006.143	10007.524	13783.501	7932.272
☒	2005	1.026	13525.18	7844.968	10050.957	4521.814
☒	2006	2.036	15908.949	9085.464	13513.279	6751.647
☐	2007	0.615	15932.346	10166.587	13162.264	7547.81
☒	2008	1.329	14662.399	9205.703	11179.314	5871.278
☒	2009	2.14	16042.875	9862.958	13469.521	7367.78
☒	2010	1.71	13303.72	8367.646	12615.822	7681.175
Target	2011	*	15054.422	8757.636	13114.903	6944.363

Exclude missing Reset Cancel OK

10. Build final model

CGMS Statistical Tool

File View Tools Help

Regression analysis

Scenario analysis

Area:

al-Djazair
al-Maghribiya
Centre (005)
Centre (010)
Centre (013)
Centre-North (003)

Crop:

wheat

Period:

Retrieve Analyst Settings

Indicators Options Output Model details

Description of model

NUTS area	al-Maghribiya
Crop	wheat
Decade	13
Included years	35
Start Year	1975
End Year	2010
Target Year	2011
Transformation of years	None
Offset for years	1965.00
Excluded years	2007
Timetrend	Linear
Included indicators	04 Water Limited Storage Organs

Summary Statistics

R-squared	44.018
-----------	--------

11. Evaluate the model

CGMS Statistical Tool

File View Tools Help

Time trend analysis
Regression analysis
 Scenario analysis

Area:
 al Diazair
 al-Maghribia
 Centre (005)
 Centre (010)
 Centre (013)
 Centre-North (003)

Crop:
 wheat

Period:
 J F M A M J J A S O N D
 I C C C C C C C C C C C
 II C C C C C C C C C C C C
 III C C C C C C C C C C C C

Retrieve Analyst Settings

Indicators Options Output **Model details**

Summary Statistics

R-squared	44.018
Adjusted R-squared	40.519
Residual Standard deviation	0.356
Root mean squared error for prediction	0.364
Mallows Cp	0.872
Maximum of VIF	1.207
Prediction for target year	1.736
Standard Error of Prediction (Mean)	0.131
Standard Error of Prediction (New)	0.379
Residual Degrees of Freedom	32.000

Regression coefficients

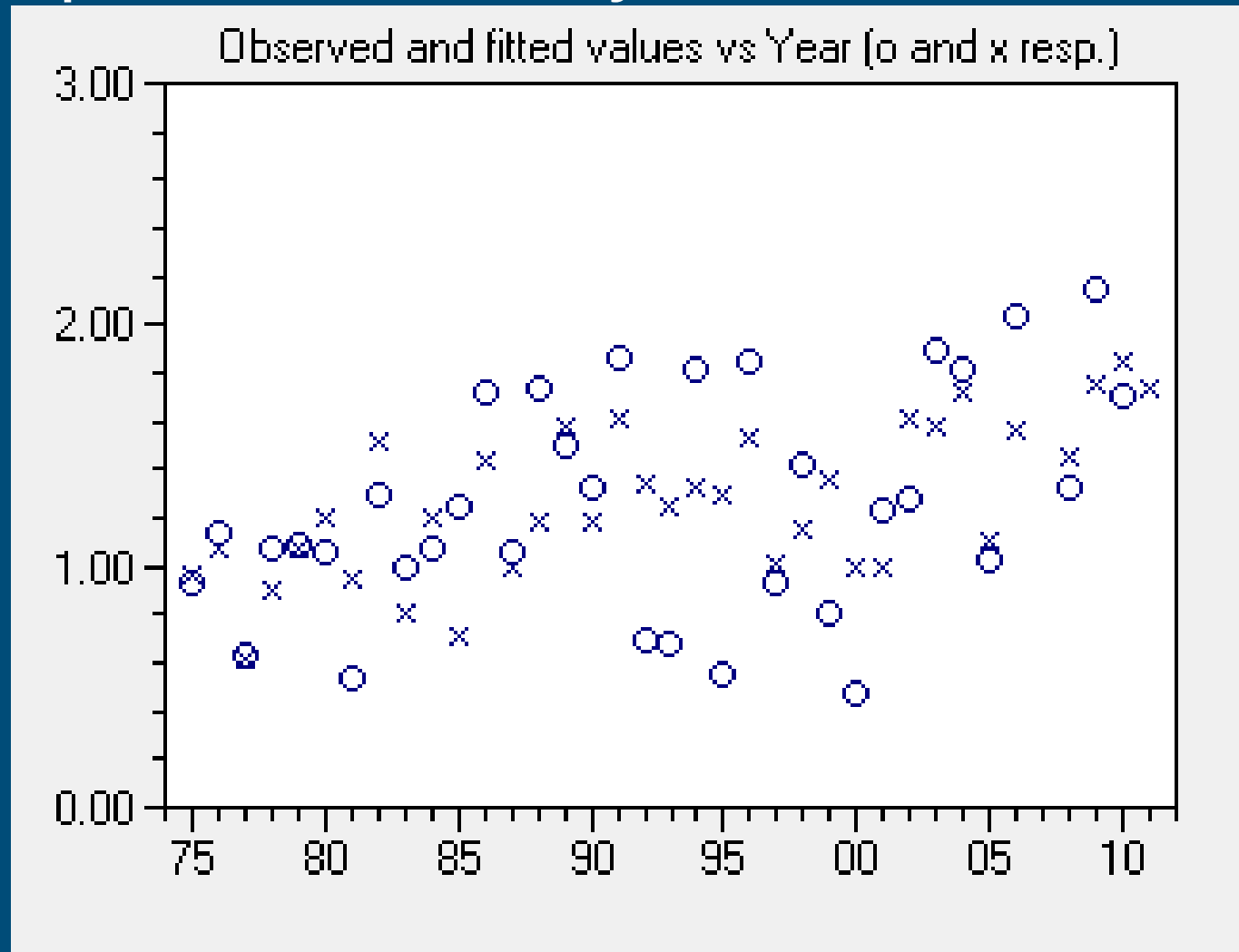
Parameter	estimate	s.e.	tvalue	t pr.	vif
Constant	-0.887	0.451	-1.966	0.06	*
Timetrend linear	0.029	6.45E-003	4.452	9.68E-005	1.21
04 Water Limited Storage Organs	1.87E-004	4.75E-005	3.947	4.07E-004	1.21

12. Apply the forecasting model

Our complete forecasting model is specified by:

1. The time trend model +
2. A linear model which can be considered as a model for yield anomalies as a function of the CGMS simulation results

13. Reported vs. fitted yields



14. Common pitfalls

- Take the time-window too large: unstable trend
- Use multiple linear regression with too many indicators (low DF): good fits but no predictive power
- Overlook collinearity
- Ignore non-statistical evidence that an outlier is indeed a bad value

- Example of scenario analysis for wheat in Portugal

Scenario analysis

CGMS Statistical Tool

File View Tools Help

Time trend analysis
Regression analysis
Scenario analysis

Area:
Monaco
Nederland
Norge
Österreich
Polska
Portugal

Crop:
wheat

Period:
J F M A M J J A S O N D
I II III

Retrieve Analyst Settings

Indicators Options **Output** Model details

Number of principal components used: 3
Percentage of variance explained: 98.820
Highest tolerated dissimilarity: 1.917

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

Yield

Year

	Model based on time trend and:	Residual std. dev.	Forecast
	nothing else	0.315	0.996
	plain residuals	0.311	0.791
	weighed residuals	0.311	0.780

Save

The CGMS statistical toolbox (CST)

Observation:

- Manual analyses is error prone
- Desire by MARS-Stat for a dedicated tool for yield forecasting

Development of CGMS Statistical Toolbox

- CST does several analyses: time trend analyses, (multiple) regression analyses and scenario analyses
- Each model is tested whether it improves prediction beyond the trend only
- Hypothesis testing for determining significance of results

Thank you

Merci

شكرا

Dankuwel

Asante sana

谢谢你