The instruction was carried out by Pawel Piotrowski, version number 1 dated 6.01.2024

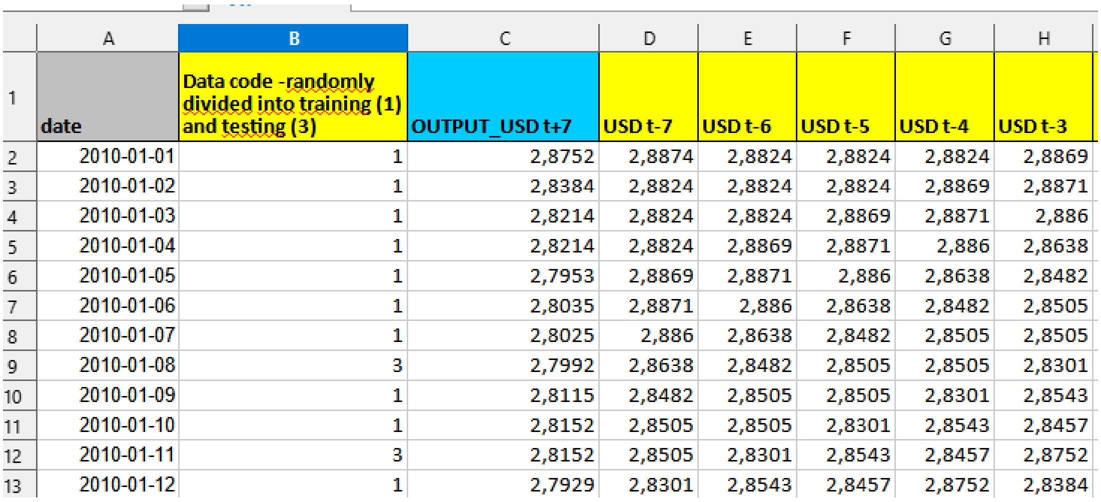
Computer exercise 3 – Short-term forecasting of US dollar exchange rate using machine learning technique - KNN

Exercise environment: Libre Office Calc (spreadsheet), Statistica program, a helper for everyone: the human brain.

The goal of the exercise is to perform data analysis (variable selection for models), forecast the USD exchange rate with a horizon of t+7 (7 forward quotations) using the KNN machine learning technique (regression problem).

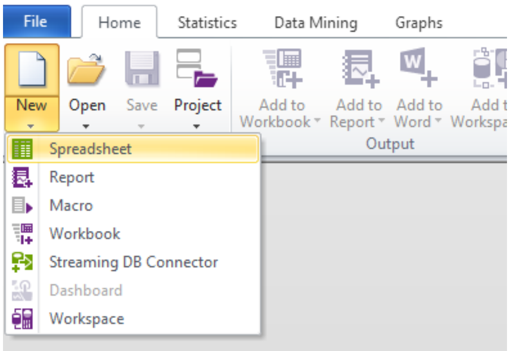
1. Statistical analysis of data

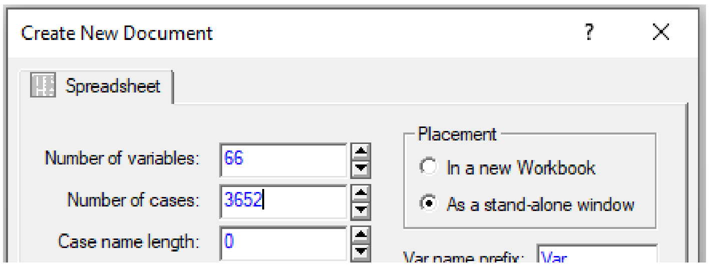
Computer_exercise_3_KNN_regression_student.xls', on the 'Input_ouput_data_by_time' tab, data is collected as a time series. Column B contains randomly assigned numbers 1 or 3 to subsequent rows while maintaining quantitative proportions (85% training data, 15% test data).



	A	B	C	D	E	F	G	H
1	date	Data code -randomly divided into training (1) and testing (3)	OUTPUT_USD t+7	USD t-7	USD t-6	USD t-5	USD t-4	USD t-3
2	2010-01-01	1	2,8752	2,8874	2,8824	2,8824	2,8824	2,8869
3	2010-01-02	1	2,8384	2,8824	2,8824	2,8824	2,8869	2,8871
4	2010-01-03	1	2,8214	2,8824	2,8824	2,8869	2,8871	2,886
5	2010-01-04	1	2,8214	2,8824	2,8869	2,8871	2,886	2,8638
6	2010-01-05	1	2,7953	2,8869	2,8871	2,886	2,8638	2,8482
7	2010-01-06	1	2,8035	2,8871	2,886	2,8638	2,8482	2,8505
8	2010-01-07	1	2,8025	2,886	2,8638	2,8482	2,8505	2,8505
9	2010-01-08	3	2,7992	2,8638	2,8482	2,8505	2,8505	2,8301
10	2010-01-09	1	2,8115	2,8482	2,8505	2,8505	2,8301	2,8543
11	2010-01-10	1	2,8152	2,8505	2,8505	2,8301	2,8543	2,8457
12	2010-01-11	3	2,8152	2,8505	2,8301	2,8543	2,8457	2,8752
13	2010-01-12	1	2,7929	2,8301	2,8543	2,8457	2,8752	2,8384

In Statistica program, select: Menu -> New -> Spreadsheet -> enter number of variables 66, cases 3652.




Create New Document

Spreadsheet

Number of variables: 66

Number of cases: 3652

Case name length: 0

Placement

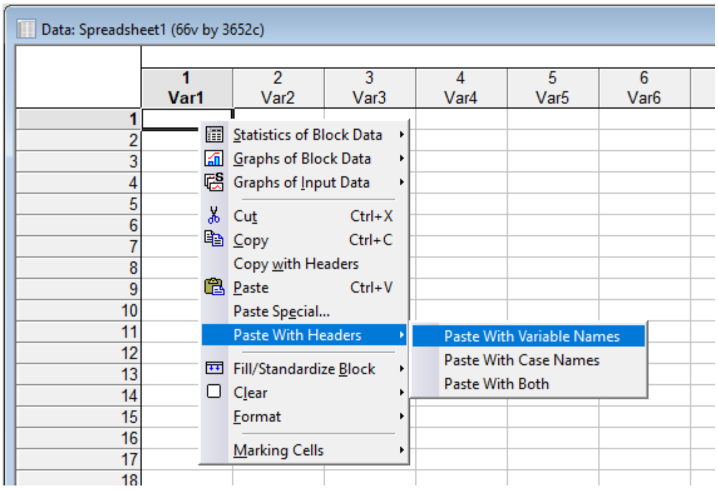
○ In a new Workbook

● As a stand-alone window

Var name prefix: Var

In LibreOffice Calc, in the 'Input_ouput_data_sorted' tab, we select the range B1:BG3653, which includes the selected data from the 'Input_output_data_sorted' tab (58 variables (including 56 potential input variables), 3652 cases).

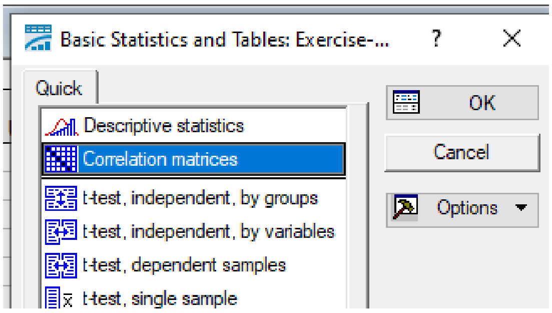
In Statistica, go to first cell, click right buton of mouse, paste the data into the sheet -> paste With Headers -> Paste With Variable Names.



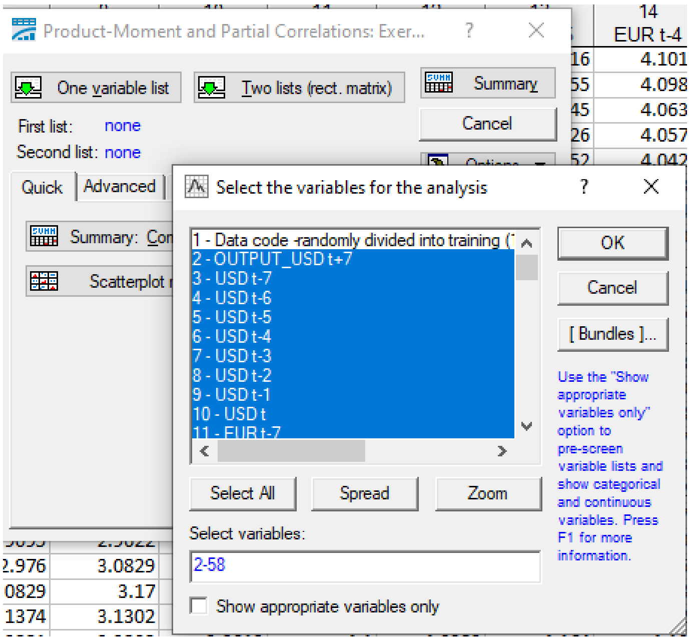
Select File->Save as and save your spreadsheet with sta extension. You can open your work by clicking in this file next time.

Variable selection based on the value of Pearson's linear correlation coefficienta

Select menu Statistics ->Basic statistics-> Correlation matrics


Basic Statistics and Tables: Exercise-...
Quick
Descriptive statistics
Correlation matrices
t-test, independent, by groups
t-test, independent, by variables
t-test, dependent samples
t-test, single sample
OK
Cancel
Options

Select->One variable list->



First list: none
Second list: none

1 - Data code -randomly divided into training (
2 - OUTPUT_USD t+7
3 - USD t-7
4 - USD t-6
5 - USD t-5
6 - USD t-4
7 - USD t-3
8 - USD t-2
9 - USD t-1
10 - USD t
11 - EUR t-7

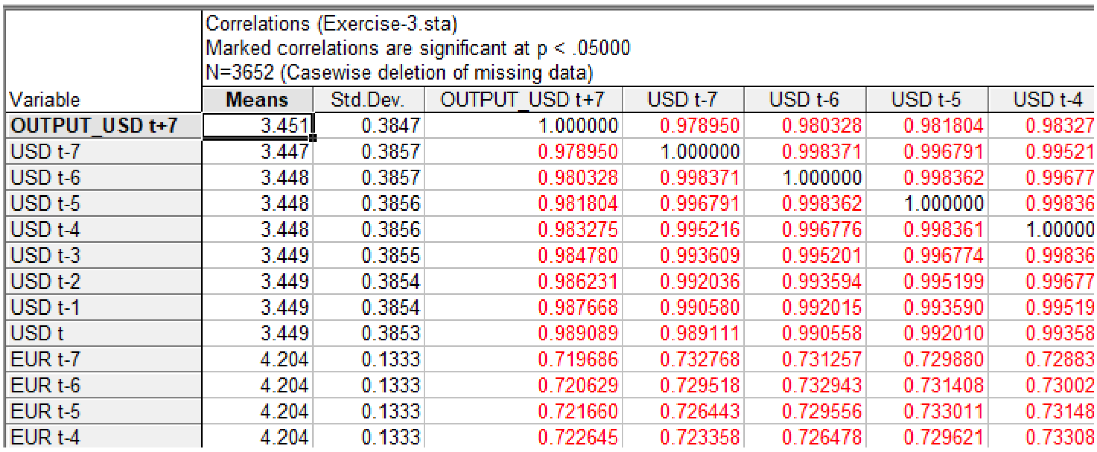
Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.

Select variables: 2-58

□ Show appropriate variables only

Enter 2-58 in the selection variable field or manually choose variables while holding down CTRL, click OK

In the Quick tab, select -> Summary: Correlation.



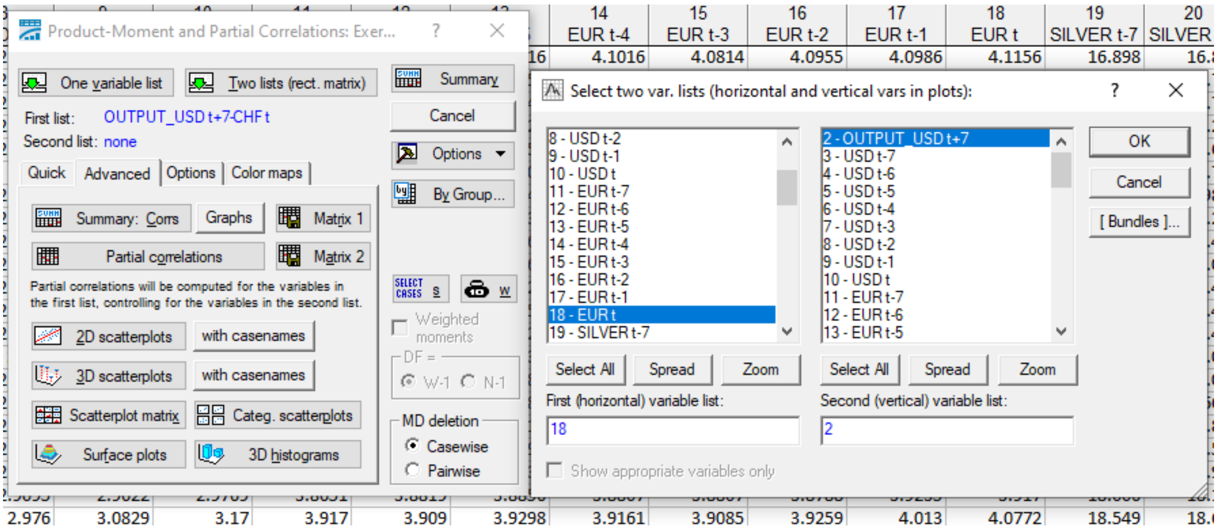
Correlations (Exercise-3.sta)
Marked correlations are significant at p < .05000
N=3652 (Casewise deletion of missing data)

Variable	Means	Std.Dev.	OUTPUT_USD t+7	USD t-7	USD t-6	USD t-5	USD t-4
OUTPUT_USD t+7	3.451	0.3847	1.000000	0.978950	0.980328	0.981804	0.98327
USD t-7	3.447	0.3857	0.978950	1.000000	0.998371	0.996791	0.99521
USD t-6	3.448	0.3857	0.980328	0.998371	1.000000	0.998362	0.99677
USD t-5	3.448	0.3856	0.981804	0.996791	0.998362	1.000000	0.99836
USD t-4	3.448	0.3856	0.983275	0.995216	0.996776	0.998361	1.00000
USD t-3	3.449	0.3855	0.984780	0.993609	0.995201	0.996774	0.99836
USD t-2	3.449	0.3854	0.986231	0.992036	0.993594	0.995199	0.99677
USD t-1	3.449	0.3854	0.987668	0.990580	0.992015	0.993590	0.99519
USD t	3.449	0.3853	0.989089	0.989111	0.990558	0.992010	0.99358
EUR t-7	4.204	0.1333	0.719686	0.732768	0.731257	0.729880	0.72883
EUR t-6	4.204	0.1333	0.720629	0.729518	0.732943	0.731408	0.73002
EUR t-5	4.204	0.1333	0.721660	0.726443	0.729556	0.733011	0.73148
EUR t-4	4.204	0.1333	0.722645	0.723358	0.726478	0.729621	0.73308

Record the results, copy them (select all, copy with headers) to a LibreOffice Calc spreadsheet in the Statistical analysis tab.

You can make a selection of input variables by analyzing the correlation coefficients of the output variable with potential input variables (statistically significant data is highlighted in red)

In the Advanced tab, select -> 2D scatterplots. Select variable (first 18 and second 2)
We can investigate what type of relationship exists (linear? non-linear?) between the output variable OUTPUT_USD t+1 and the selected potential input variable, such as EUR t.



First list: OUTPUT_USD t+7-CHF t

Second list: none

Partial correlations will be computed for the variables in the first list, controlling for the variables in the second list.

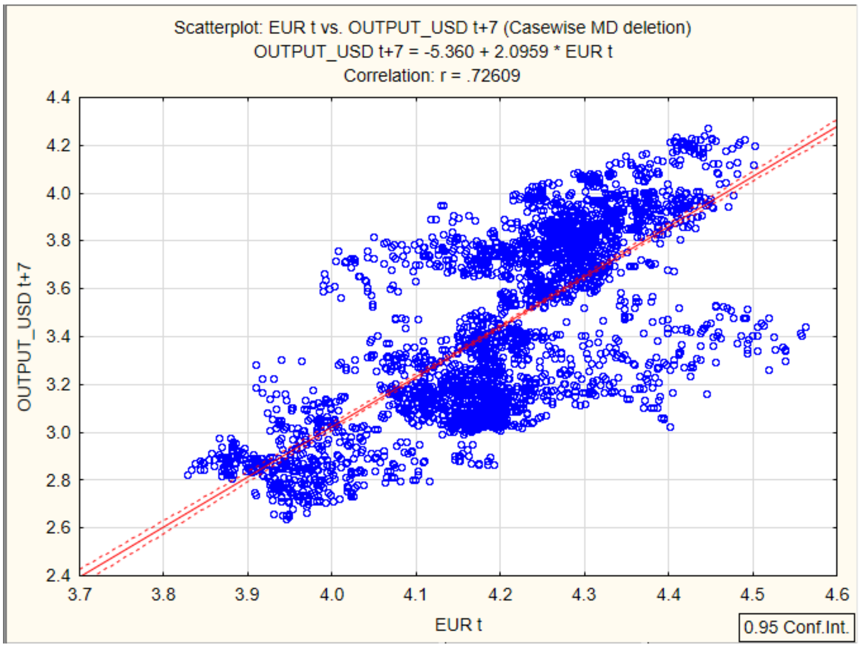
MD deletion: ● Casewise ○ Pairwise

First (horizontal) variable list: 18

Second (vertical) variable list: 2

14 EUR t-4	15 EUR t-3	16 EUR t-2	17 EUR t-1	18 EUR t	19 SILVER t-7	20 SILVER
4.1016	4.0814	4.0955	4.0986	4.1156	16.898	16.

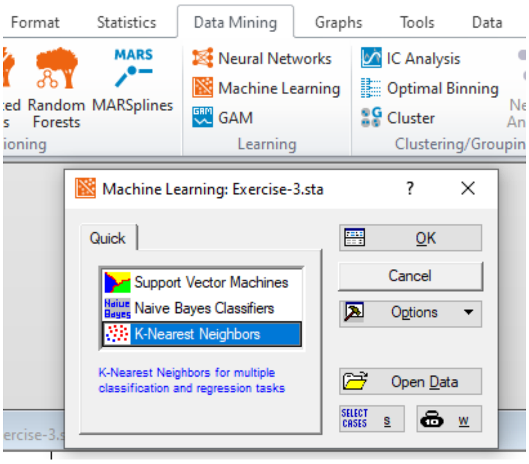
2.976	3.0829	3.17	3.917	3.909	3.9298	3.9161	3.9085	3.9259	4.013	4.0772	18.549	18.



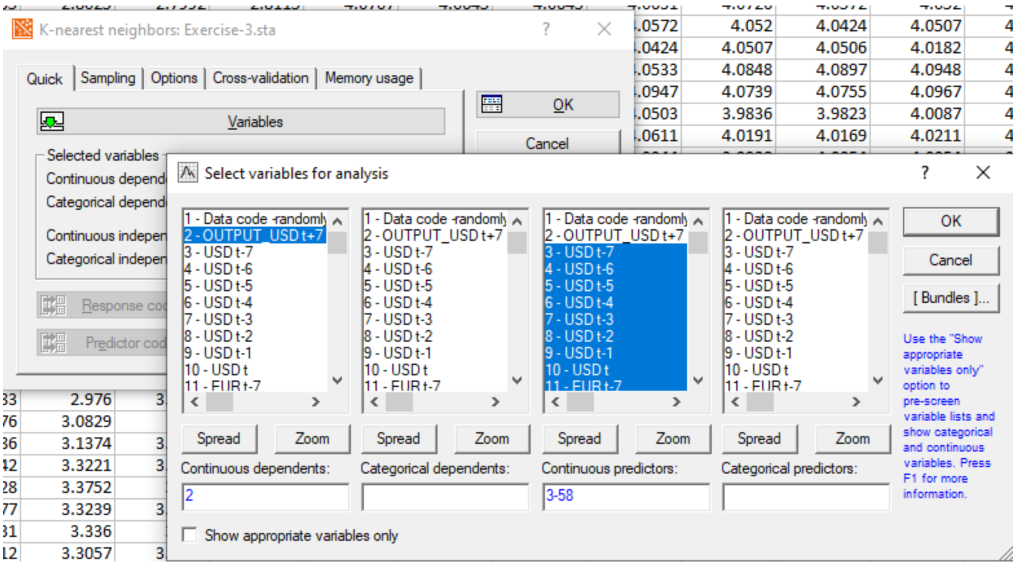
You can click right buton of mouse on the graph area and select save as – select png format of figure – for report preparation.

3. Forecasts using machine learning technique - KNN regression for a horizon of t+7

Select menu -> Data Mining -> Machine learning->select „K-Nearest Neighbors"



In Tab Quick select buton Variables and select output – number 2 (continous dependents) next select inputs (inputs numbered from 3 to 48) - while holding the CTRL key, you can select or enter numerically in the field.



Continuous dependents: 2

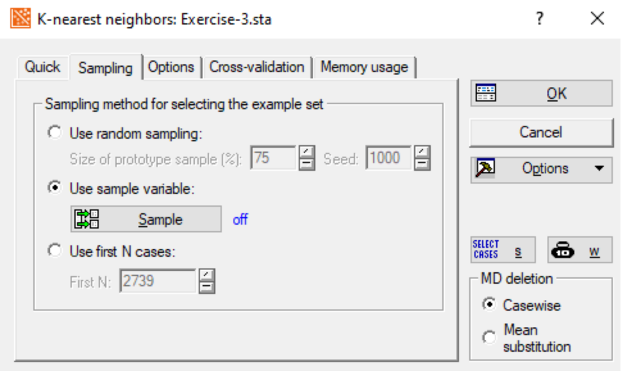
Categorical dependents:

Continuous predictors: 3-58

Categorical predictors:

□ Show appropriate variables only

In next Tab „sampling”, select „Use sample variable”, select buton „sample”



Sampling method for selecting the example set

○ Use random sampling:

Size of prototype sample (%): 75 Seed: 1000

● Use sample variable:

○ Use first N cases:

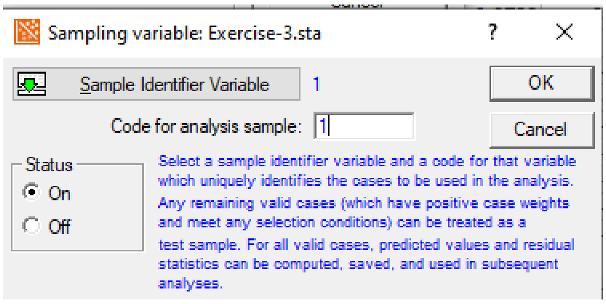
First N: 2739

MD deletion

● Casewise

○ Mean substitution

Next select Status – „on”, select „Sample Identifier Variable”, next select „Data code - randomly divided into training (1) and testing (3)”, select OK. Write 1 in field „Code for analysis sample (1 -is the range of training data for KNN).



Sampling variable: Exercise-3.sta

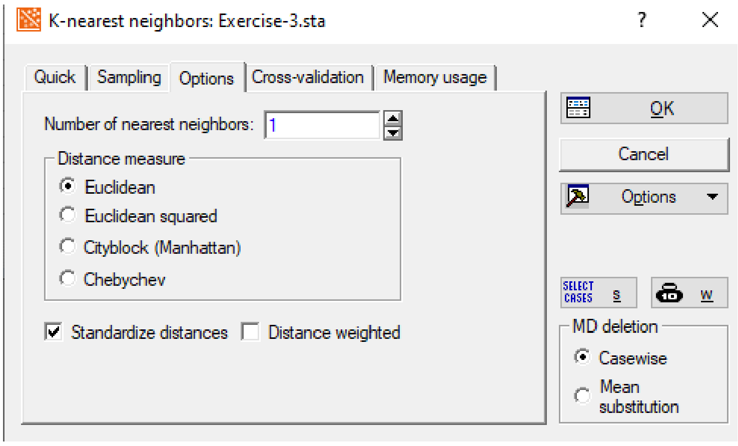
Sample Identifier Variable 1

Code for analysis sample: 1

Status
● On
○ Off

Select a sample identifier variable and a code for that variable which uniquely identifies the cases to be used in the analysis. Any remaining valid cases (which have positive case weights and meet any selection conditions) can be treated as a test sample. For all valid cases, predicted values and residual statistics can be computed, saved, and used in subsequent analyses.

Next select Tab „Options”, select Distance measure – Euclidean, select – Standardize distances, number of nearset neighbors – select – 1 because the number of nearest neighbors will be determined automatically (grid search method).



K-nearest neighbors: Exercise-3.sta

Quick | Sampling | Options | Cross-validation | Memory usage

Number of nearest neighbors: 1

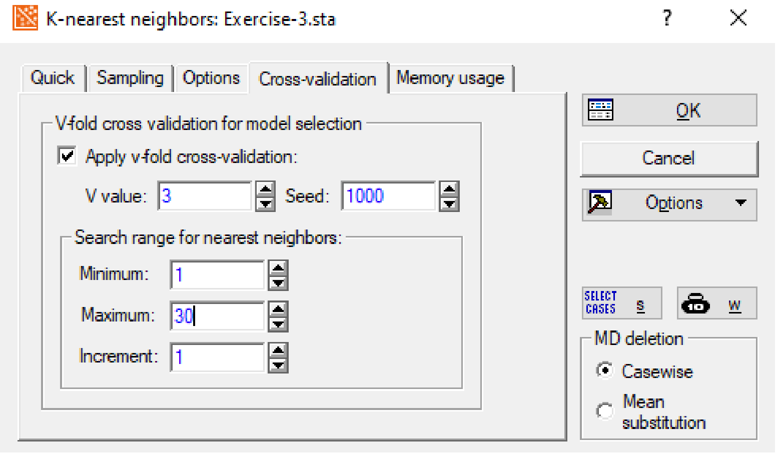
Distance measure
● Euclidean
○ Euclidean squared
○ Cityblock (Manhattan)
○ Chebychev

☑ Standardize distances ☐ Distance weighted

MD deletion
● Casewise
○ Mean substitution

Next, select Tab „Cross-validation” select „Apply v-fold cross-validation, select „V value” – 3, select „Maximum” – 30.

Note: A detailed description of the algorithm's operation can be found by clicking on the "?" in the upper right corner.



K-nearest neighbors: Exercise-3.sta

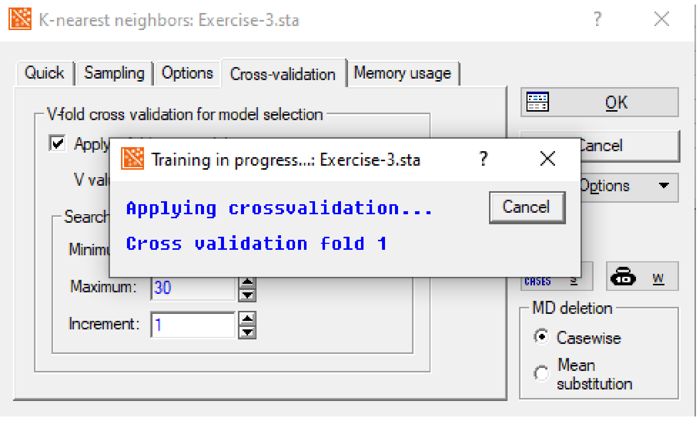
Quick | Sampling | Options | Cross-validation | Memory usage

V-fold cross validation for model selection
☑ Apply v-fold cross-validation:
V value: 3 Seed: 1000

Search range for nearest neighbors:
Minimum: 1
Maximum: 30
Increment: 1

MD deletion
● Casewise
○ Mean substitution

Select OK, you can see training in progres..



V-fold cross validation for model selection

☑ Apply

Maximum: 30

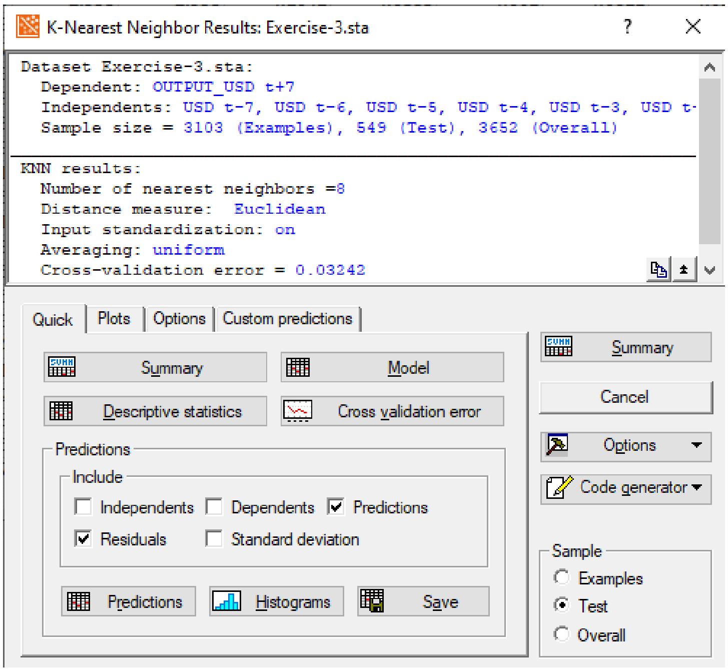
Increment: 1

● Casewise

○ Mean substitution

```
Applying crossvalidation...
Cross validation fold 1
```

In „Quick” tab select „Sample” – Test, you can see the best number of nearest neighbors in KNN results window.



```
Dataset Exercise-3.sta:
   Dependent: OUTPUT_USD t+7
   Independents: USD t-7, USD t-6, USD t-5, USD t-4, USD t-3, USD t-
   Sample size = 3103 (Examples), 549 (Test), 3652 (Overall)

KNN results:
   Number of nearest neighbors =8
   Distance measure:  Euclidean
   Input standardization: on
   Averaging: uniform
   Cross-validation error = 0.03242
```

☐ Independents ☐ Dependents ☑ Predictions

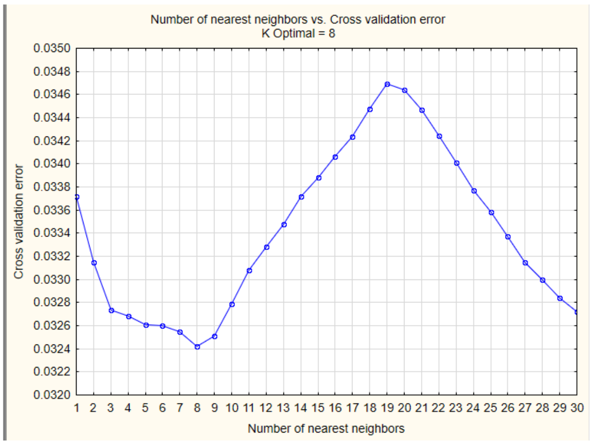
☑ Residuals ☐ Standard deviation

○ Examples

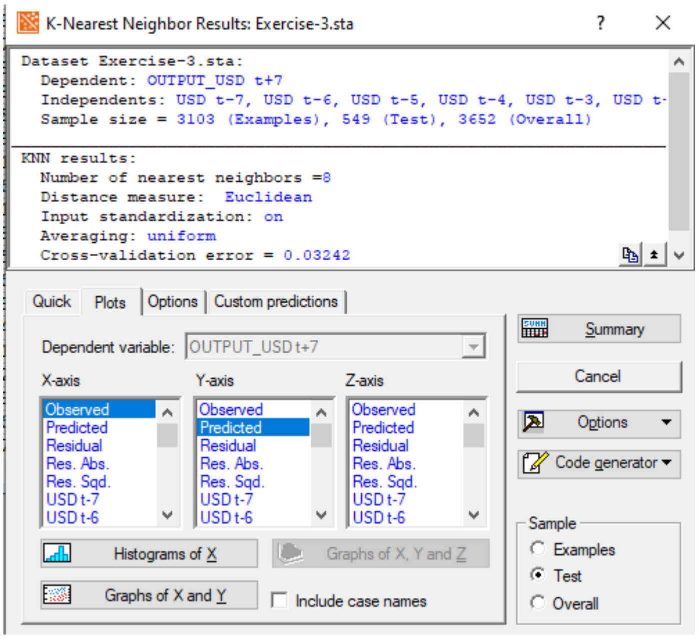
● Test

○ Overall

Select buton „Cross validation error”. You can click right buton of mouse on the graph area and select save as – select png format of figure – for report preparation.



Next, select tab „Plots”, select buton „Graphs of X and Y”



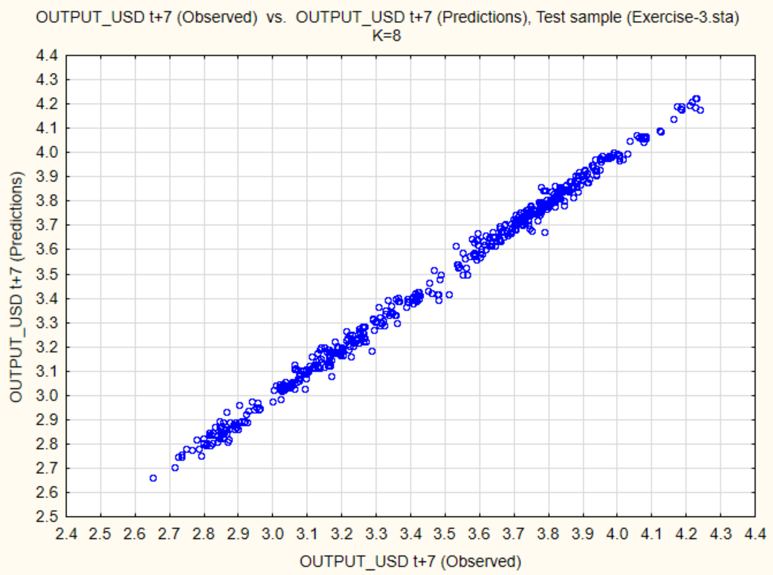
```
Dataset Exercise-3.sta:
  Dependent: OUTPUT_USD t+7
  Independents: USD t-7, USD t-6, USD t-5, USD t-4, USD t-3, USD t-
  Sample size = 3103 (Examples), 549 (Test), 3652 (Overall)

KNN results:
  Number of nearest neighbors =8
  Distance measure:  Euclidean
  Input standardization: on
  Averaging: uniform
  Cross-validation error = 0.03242
```

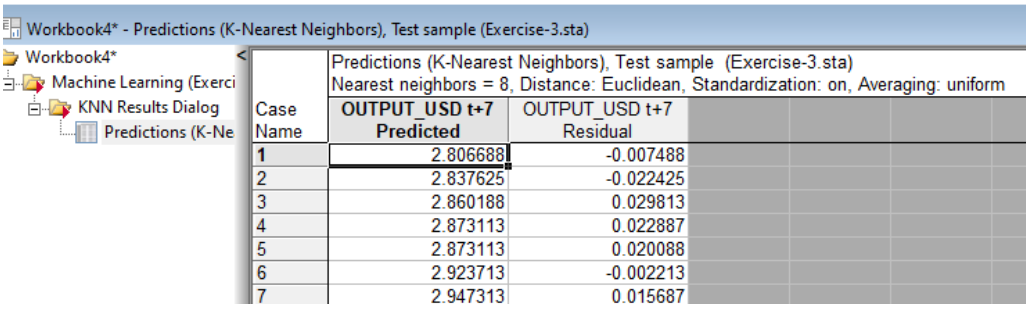
Dependent variable: OUTPUT_USD t+7

Sample: ○ Examples ● Test ○ Overall

You can click right buton of mouse on the graph area and select save as – select png format of figure – for report preparation. The chart illustrates the relationship between actual values and predicted values - in an ideal situation (no prediction errors), the points would form a straight line at a 45-degree angle.



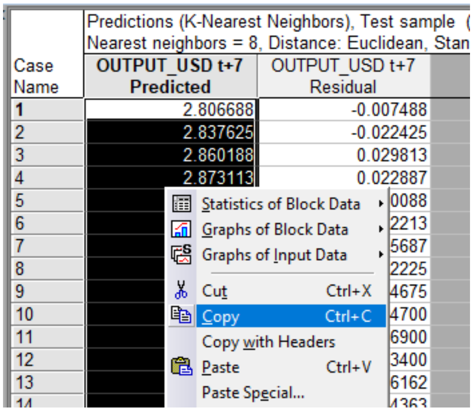
Change Tab into „Quick” select buton „Predictions”. Select all data in column Predicted.


Workbook4* - Predictions (K-Nearest Neighbors), Test sample (Exercise-3.sta)
Workbook4*
Machine Learning (Exerci
KNN Results Dialog
Predictions (K-Ne

Predictions (K-Nearest Neighbors), Test sample (Exercise-3.sta)
Nearest neighbors = 8, Distance: Euclidean, Standardization: on, Averaging: uniform

Case Name	OUTPUT_USD t+7 Predicted	OUTPUT_USD t+7 Residual
1	2.806688	-0.007488
2	2.837625	-0.022425
3	2.860188	0.029813
4	2.873113	0.022887
5	2.873113	0.020088
6	2.923713	-0.002213
7	2.947313	0.015687

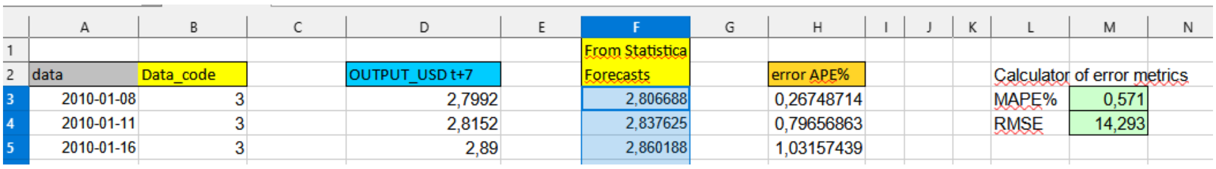
Click right buton of mouse and select Copy.



Predictions (K-Nearest Neighbors), Test sample (
Nearest neighbors = 8, Distance: Euclidean, Stan

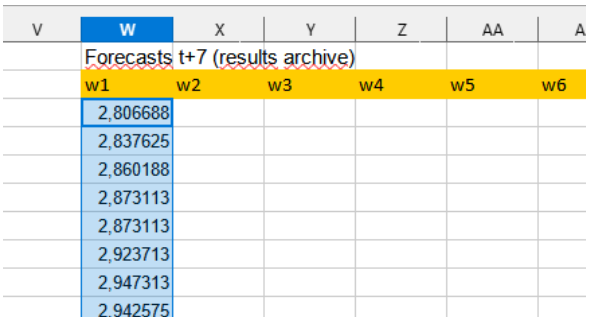
Case Name	OUTPUT_USD t+7 Predicted	OUTPUT_USD t+7 Residual
1	2.806688	-0.007488
2	2.837625	-0.022425
3	2.860188	0.029813
4	2.873113	0.022887
5		0088
6		2213
7		5687
8		2225
9		4675
10		4700
11		6900
12		3400
13		6162
14		4363

Move to LibreOffice Cal into Tab „KNN_results” and click cell F3 and click right buton of mouse and select „Paste” You can see in Calculator of errors metrics results” MAPE% error and RMSE error (main error to minimise).



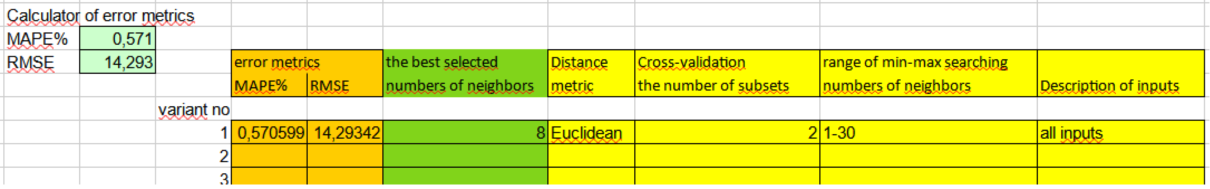
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1						From Statistica								
2	data	Data_code		OUTPUT_USD t+7		Forecasts		error APE%				Calculator of error metrics		
3	2010-01-08	3		2,7992		2,806688		0,26748714				MAPE%	0,571	
4	2010-01-11	3		2,8152		2,837625		0,79656863				RMSE	14,293	
5	2010-01-16	3		2,89		2,860188		1,03157439						

Also click cell W3 and click right buton of mouse and select „Paste” (result archive)



V	W	X	Y	Z	AA	A
	Forecasts t+7 (results archive)					
	w1	w2	w3	w4	w5	w6
	2,806688					
	2,837625					
	2,860188					
	2,873113					
	2,873113					
	2,923713					
	2,947313					
	2,942575					

Save the data for the tested variant in LibreOffice Calc.



Calculator of error metrics									
MAPE%	0,571								
RMSE	14,293		error metrics		the best selected	Distance	Cross-validation	range of min-max searching	
			MAPE%	RMSE	numbers of neighbors	metric	the number of subsets	numbers of neighbors	Description of inputs
		variant no							
		1	0,570599	14,29342	8	Euclidean	2	1-30	all inputs
		2							
		3							

Independent work:

You are seeking the best result (minimizing RMSE error) by manipulating the distance measure, the range of the minimum-maximum number of nearest neighbors (if necessary), and the number of subsets in cross-validation. We can also change the list of input variables. At least 4 study variants should be performed.

Write a research report in pdf file

Suggested points: brief introduction to KNN, presentation of results, final conclusions