

1 Data import and preparation

This part of the document deals with the data preparation of the provided cooper wire data before the data analysis.

1.1 Data import

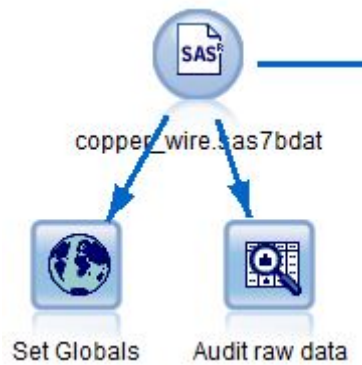


Abbildung 1: Data import in Stream

The data is imported via the node *SAS file*. The node *Set Globals* is used for setting the audited data results of the raw imported data as global values, which get used later on for the data preparation. The node *Data Audit* is used for analyzing the raw data.

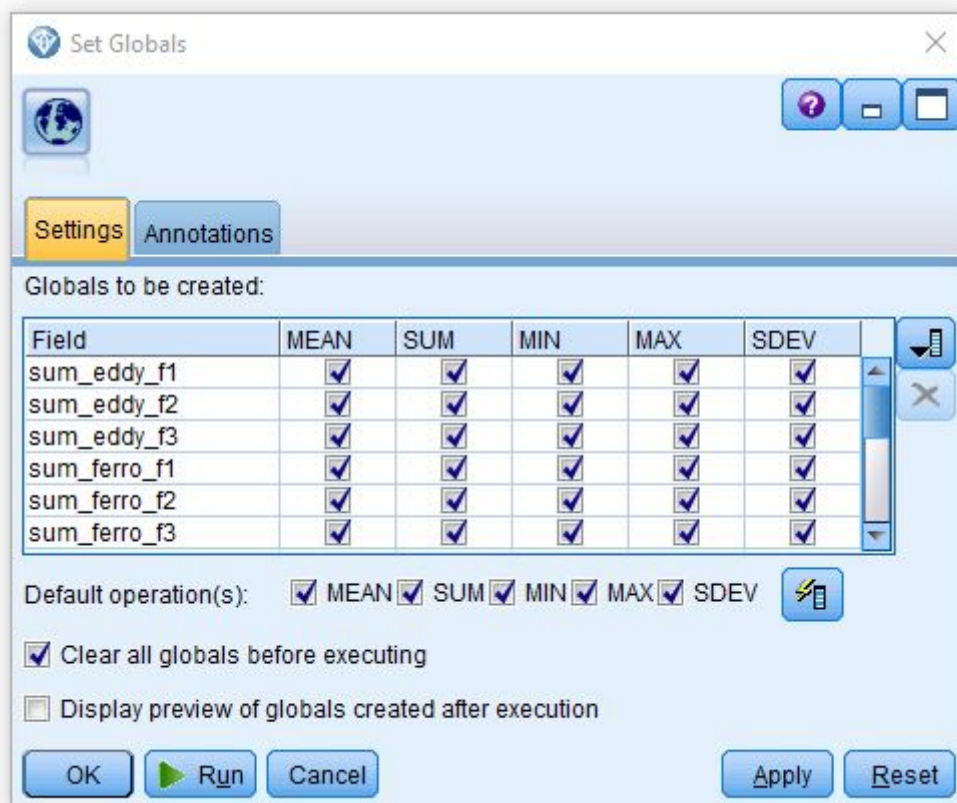
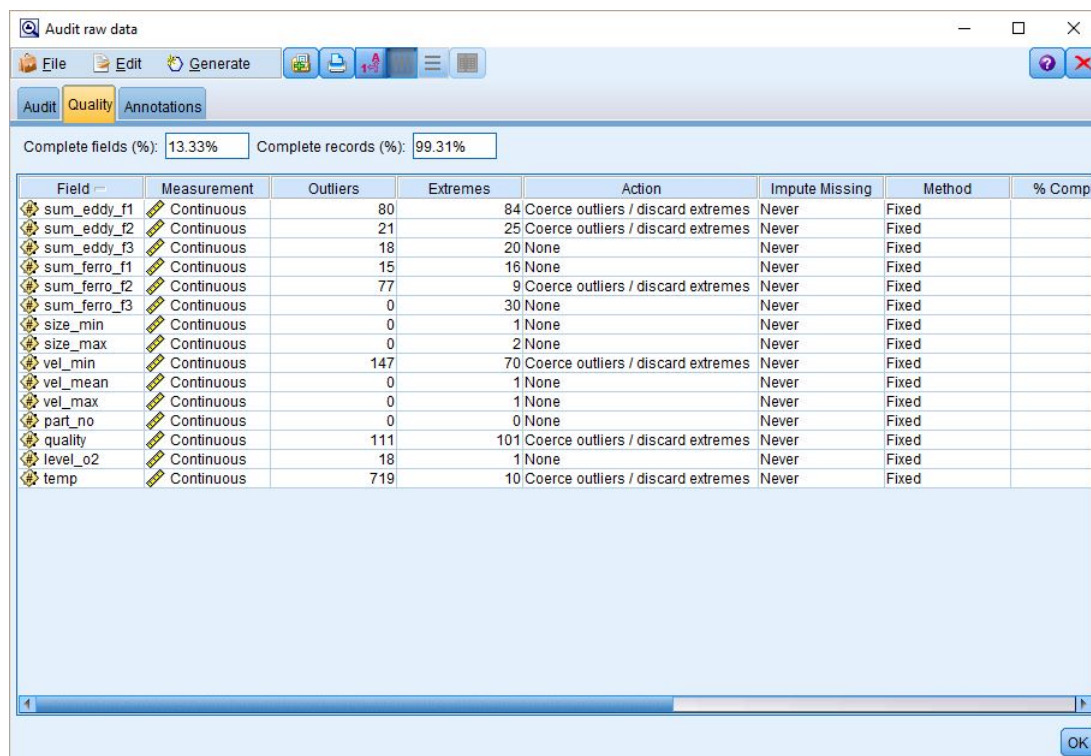


Abbildung 2: Set audit results as global values in the stream

1.2 Data preparation

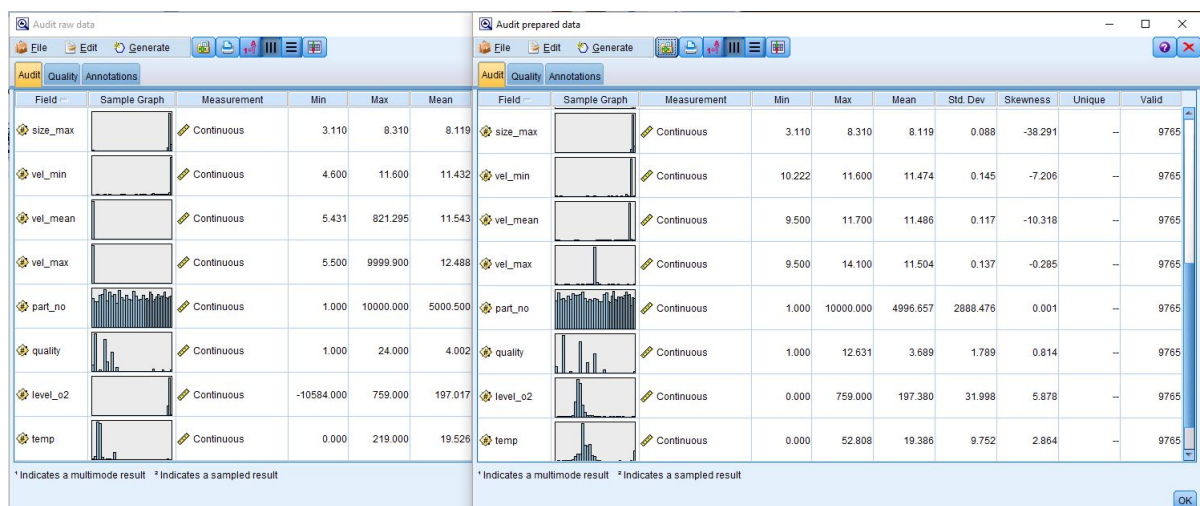
The outliers and extremes were determined during the audit of the raw data.



The screenshot shows the 'Audit raw data' window with the 'Quality' tab selected. It displays a table with columns: Field, Measurement, Outliers, Extremes, Action, Impute Missing, Method, and % Compl. The table lists 14 fields with their respective measurement types, outlier counts, extreme counts, and actions taken.

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Compl
sum_eddy_f1	Continuous	80	84	Coerce outliers / discard extremes	Never	Fixed	
sum_eddy_f2	Continuous	21	25	Coerce outliers / discard extremes	Never	Fixed	
sum_eddy_f3	Continuous	18	20	None	Never	Fixed	
sum_ferro_f1	Continuous	15	16	None	Never	Fixed	
sum_ferro_f2	Continuous	77	9	Coerce outliers / discard extremes	Never	Fixed	
sum_ferro_f3	Continuous	0	30	None	Never	Fixed	
size_min	Continuous	0	1	None	Never	Fixed	
size_max	Continuous	0	2	None	Never	Fixed	
vel_min	Continuous	147	70	Coerce outliers / discard extremes	Never	Fixed	
vel_mean	Continuous	0	1	None	Never	Fixed	
vel_max	Continuous	0	1	None	Never	Fixed	
part_no	Continuous	0	0	None	Never	Fixed	
quality	Continuous	111	101	Coerce outliers / discard extremes	Never	Fixed	
level_o2	Continuous	18	1	None	Never	Fixed	
temp	Continuous	719	10	Coerce outliers / discard extremes	Never	Fixed	

Abbildung 3: Audit of the raw data



The screenshot shows the 'Audit prepared data' window with the 'Quality' tab selected. It displays a table with columns: Field, Sample Graph, Measurement, Min, Max, Mean, Std. Dev, Skewness, Unique, and Valid. The table lists 14 fields with their respective measurement types, statistical values, and unique counts. Each field has a small sample graph icon next to it.

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
size_max		Continuous	3.110	8.310	8.119	0.088	-38.291	9765	
vel_min		Continuous	4.600	11.600	11.432	0.145	-7.206	9765	
vel_mean		Continuous	5.431	821.295	11.543	0.117	-10.318	9765	
vel_max		Continuous	5.500	9999.900	12.488	0.137	-0.285	9765	
part_no		Continuous	1.000	10000.000	5000.500	2888.476	0.001	9765	
quality		Continuous	1.000	24.000	4.002	1.789	0.814	9765	
level_o2		Continuous	-10584.000	759.000	197.017	31.998	5.878	9765	
temp		Continuous	0.000	219.000	19.526	9.752	2.884	9765	

Abbildung 4: Audit of the raw data

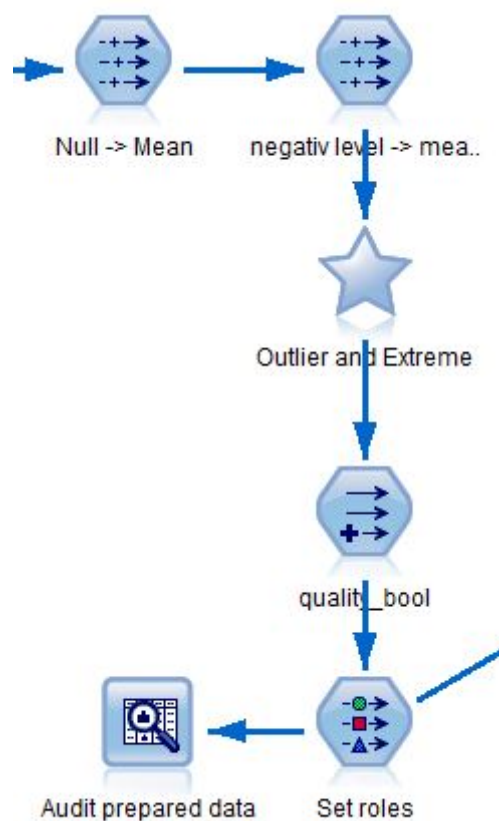


Abbildung 5: Flow of data preparation tasks

This flow prepares the data for the later analysis. The following tasks are performed:

- Null values will be replaced with the mean value set by the *Set Global* node
- The negative value of the field *temp* will be replaced with the global mean of this field
- The outliers and extremes will be handled as you can see in image 3
- A new field will be created *quality_bool* which represents the quality state good or false
- The fields which are not considered to be relevant will be set as ignored and the field *quality_bool* will be set as the target field for the further analysis

1.3 Predictive Model

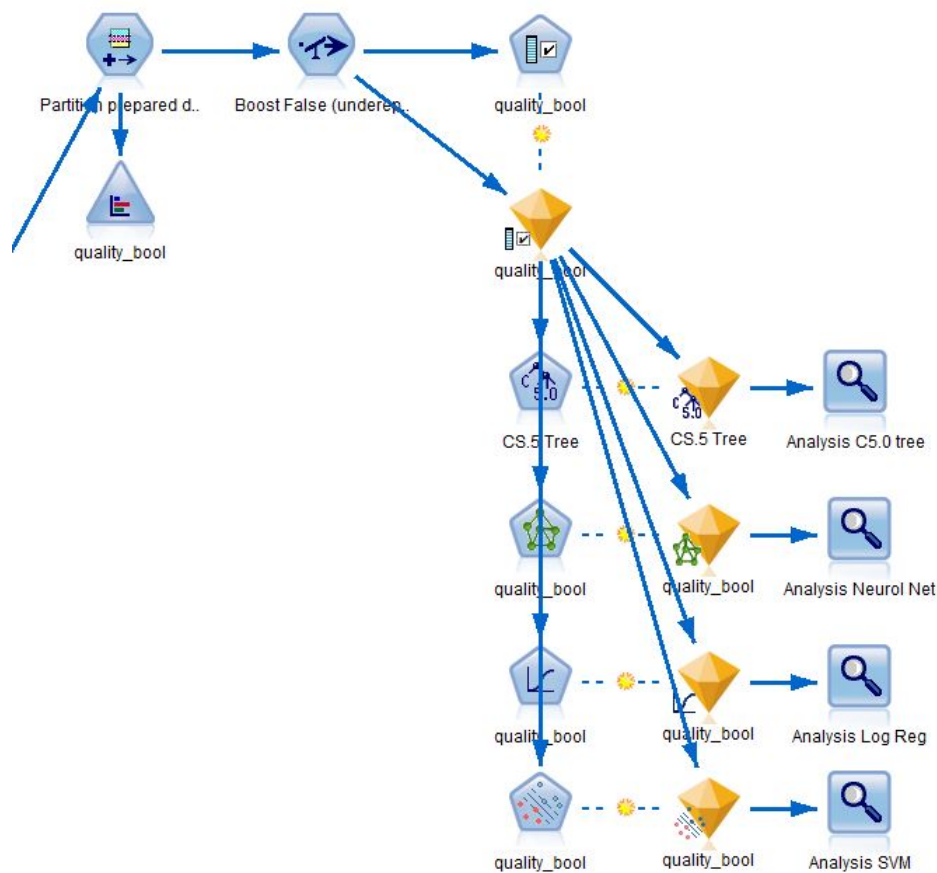


Abbildung 6: Flow of the predictive model

The flow of the predictive part of the model uses the following nodes:

- *C5.0* (The decision tree)
- *Neural Net*
- *Logistic* (Logistic Regression)
- *SVM*

Each modeling node gets analyzed with the node *Analyze*.