Project information

Project title	Coordinating Optimisation of Complex Industrial Processes
Project acronym	COCOP
Project call	H2020-SPIRE-2016
Grant number	723661
Project duration	1.10.2016-31.3.2020 (42 months)

Document information

Deliverable number	D3.3
Deliverable title	Pre-processing tools for collected data
Version	1.0
Dissemination level	Pub
Work package	WP3
Authors	BFI
Contributing partners	VTT, TUT , TEC, SID, MSI
Delivery date	26.09.2018
Planned delivery month	M24
Keywords	COCOP, data pre-processing, outlier detection, filtering, data transformation, sub-sampling

This project has received European Union's Horizon 2020 research and innovation funding under grant agreement No 723661.

Version history

Version	Description	Organisation	Date
0.1	Initial document	BFI	06.06.2018
0.2	Content added	TEC	04.08.2018
0.3	Content added	TUDO	05.09.2018
0.4	Content added	BFI	05.09.2018
0.5	First review	DSM	06.09.2018
0.6	Review remarks added	BFI	12.09.2018
0.7	Internal review	VTT	13.09.2018
0.8	Review remarks added	BFI	13.08.2018
0.9	Review remarks added	BFI, DSM, TUDO	17.09.2018
1.0	Final version for upload	BFI	17.09.2018

Executive Summary

Using repositories of data measured from demanding industrial environments raises needs for data pre-processing actions that are evaluated by applying exploratory analysis methods. In addition, suggestions and requests from operators and process experts are considered wherever needed and possible. Those requirements initiate the investigations in methods such as re-sampling, filtering, identification and removal of noisy data, and replacement by interpolated/extrapolated or modelled data.

Delays will be compensated from time-series and dynamic data, taking into account that non-stationary phenomenon that can make those dynamics and delays evolve over time. The total wall-clock time during which a production process takes place from adding the feedstock at t=0 until the product comes out of the process usually take several hours up to more than 24 hours for the entire process. This mean that data which are collected during the process have to be given a time-shift with respect to t=0 in accordance with where one is in the process.

Abbreviations

Abbreviation	Full name
KDD	Knowledge Discovery in Databases
EDA	Exploratory Data Analysis
MAD	Median absolute deviation
LOF	Local outlier factor
FFT	Fast Fourier Transformation

Table of Contents

1 Introduction 1

2 Data analysis / Data Mining 1

2.1 Data from Copper use case 2

2.2 Data from Steel use case 2

3 Description of methods for data pre-processing 3

3.1 Outlier detection 3

3.1.1 Median absolute deviation (MAD) 4

3.1.2 Local outlier factor (LOF) 4

3.1.3 Clustering (k-means) 4

3.2 Filtering / noise reduction 5

3.2.1 Frequency based filters 5

3.2.2 Savitzky-Golay 5

3.2.3 Wiener filter 6

3.2.4 Gaussian filter 6

3.3 Sub-sampling / interpolation 6

3.3.1 Linear interpolation 6

3.3.2 Cubic spline interpolation 7

3.4 Handling of time delays 7

4 Pre-processing tools from a social innovation perspective 9

5 Conclusion 11

6 References 12

1 Introduction

The need for critical analysis of the available data must be emphasised. A model cannot be better than the data used for its estimation. Exploratory data analysis (EDA) [1] methods were used to evaluate the needs for data pre-processing. Methods such as re-sampling, filtering, identification and removal of noisy data, and replacement by interpolated/extrapolated or modelled data were investigated. Delays were compensated from time-series taking into account that non-stationary phenomena can make dynamics and delays evolve over time. Scope of the analysis and selection of techniques to address are e.g. frequency domain algorithms, linear or non-linear regression, data scaling and normalization, outliers filtering, missing data and signal reconstruction.

2 Data analysis / Data Mining

“Data mining is the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules.” [2] While there are many other accepted definitions of data mining, this one captures the notion that data miners are searching for meaningful patterns in large quantities of data. The implied goal of such an effort is the use of these meaningful patterns to improve production processes or product quality. Historically the finding of useful patterns in data has been referred to as knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing in addition to data mining. In recent years, the field has settled on using the term data mining to describe these activities. [3] Statisticians have commonly used this term to refer to the patterns in data that are discovered through multivariate regression analyses and other statistical techniques.

As the evolution of data mining has matured, it is widely accepted to be a single phase in a larger life cycle known as Knowledge Discovery in Databases (KDD). The term KDD was coined in 1989 to refer to the broad process of finding knowledge in data stores. [4] The field of KDD is particularly focused on the activities leading up to the actual data analysis and including the evaluation and deployment of results. KDD nominally encompasses the following activities (see Figure 1):

1. Data Selection – The goal of this phase is the extraction from a larger data store of only the data that is relevant to the data mining analysis. This data extraction helps to streamline and speed up the process.
2. Data Pre-processing – This phase is concerned with data cleansing (also known as data curation) and preparation tasks that are necessary to ensure correct results. Strategies how to handle missing values in the data (e.g. interpolation, fill with mean values), ensuring that coded values have a uniform meaning and ensuring that no spurious data values exist are typical actions that occur during this phase.

3. Data Transformation – This phase of the lifecycle is aimed at converting the data into a two-dimensional table and eliminating unwanted or highly correlated fields so the results are valid.
4. Data Mining – The goal of the data-mining phase is to analyse the data by an appropriate set of algorithms in order to discover meaningful patterns and rules and produce predictive models. This is the core element of the KDD cycle.
5. Interpretation and Evaluation – While data mining algorithms have the potential to produce an unlimited number of patterns hidden in the data, many of these may not be meaningful or useful. This final phase is aimed at selecting those models that are valid and useful for making future business or technical decisions.

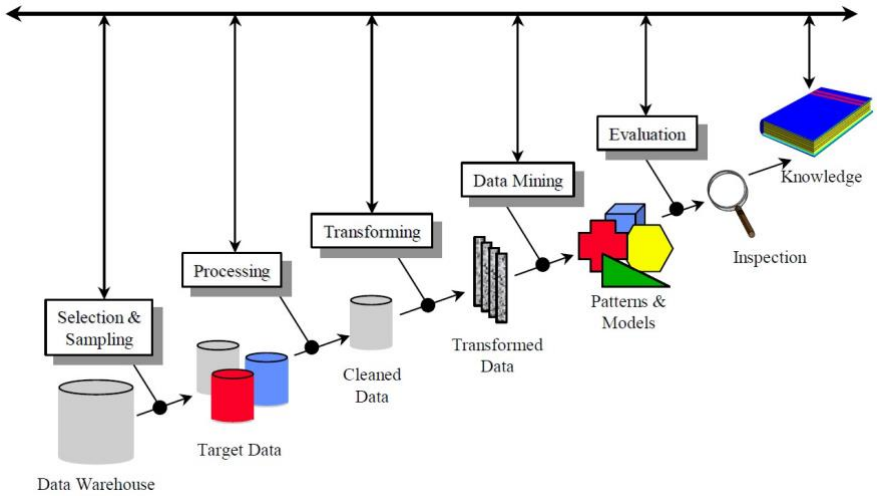

Figure 1: The Traditional KDD Paradigm [5]

This deliverable covers the steps two and three of the abovementioned activities. The steps four and five will be addressed in deliverable D3.6

2.1 Data from Copper use case

For the development of the copper specific models and optimisation strategies data from Boliden's production site Harjavalta, Finland are available. The data from one month production with a 10 second sample rate are the basis for the development. A detailed description of the data is given in deliverable D3.2.

2.2 Data from Steel use case

For the development of the steel pilot case data based models, SIDENOR process data from their production site in Basauri (Bizkaia), Spain, are available. Due to the revamping of the continuous casting, re-heating furnace and hot rolling installations, initially data from the last 5 months are being used. The description of these data is included in deliverable D3.2

3 Description of methods for data pre-processing

The field of data processing covers a wide range of methods and functionalities. This deliverable covers only a small part of the available functions that are used in COCOP project. It is not the claim to have a comprehensive explanation of all possible approaches.

3.1 Outlier detection

Data coming from industrial processes are affected by different disturbances. One of these disturbances are outliers. These are single data values in time series that does not fit into the information behind the data. Figure 2 gives an example of univariate outliers. They can be manually detected knowing e.g. the measured physical value like temperature or speed. Both cannot have such gradients. So one possible technique is the manual removal of outliers by a visual analysis of the data.

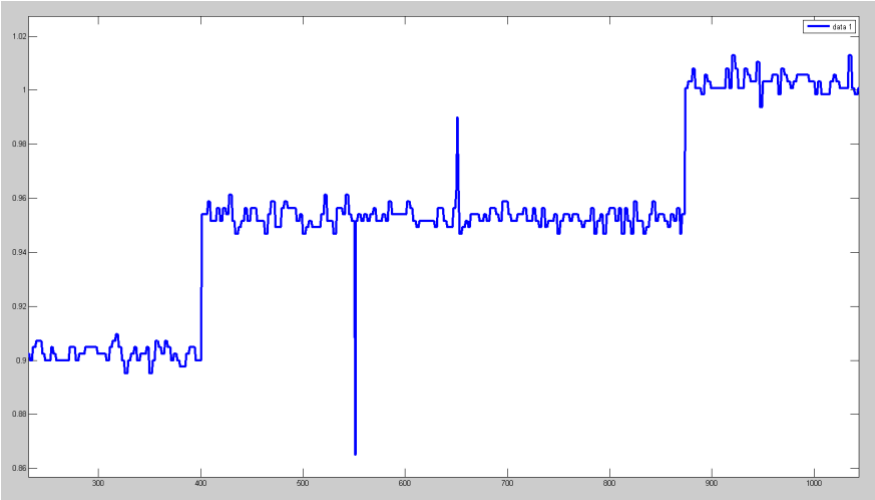

Figure 2: Univariate outliers

More complex is the detection of outliers that belong to a combination of data that leads to a higher dimensional data space that cannot be handled by human visual inspection at least in time. In Figure 3 below, an exemplary outlier in a 2D case is shown. Here the outlier (red dot) is visible by the combination of two time series. For the cases, having more than two signals a suitable method has to be applied to efficiently detect the outliers. In the following chapters, the methods that were used in COCOP are described.

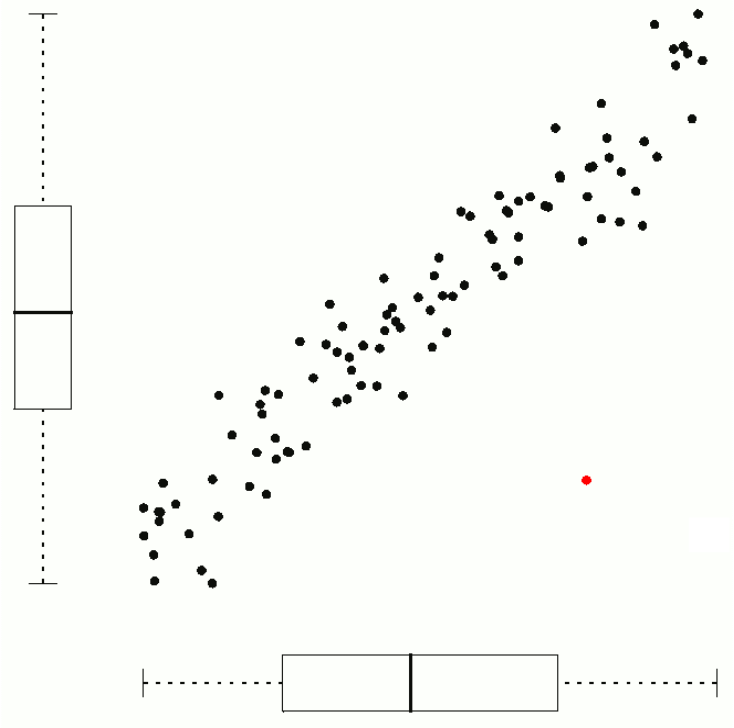

Figure 3: Example of higher dimensional outliers, here as 2D outlier

3.1.1 Median absolute deviation (MAD)

Typical approaches to outlier detection are based on the statistics of the historical data [6]. The simplest approach is the 3σ outlier detection algorithm that is based on univariate observations of the variable distributions. This method labels all data samples out of the range $\mu(x) \pm 3\sigma(x)$, where $\mu(x)$ is the mean value and $\sigma(x)$ the standard deviation of the variable x, as outliers. More robust version of this approach is the Hampel identifier that in contrast to the 3σ method uses more outlier resistant median and median absolute deviation from median (MAD) values to calculate the limits.

3.1.2 Local outlier factor (LOF)

The LOF algorithm [7] is an unsupervised outlier detection method which computes the local density deviation of a given data point with respect to its neighbours. It considers as outliers those samples that have a substantially lower density than their neighbours have. It is local in that the anomaly score depends on how isolated the object is with respect to the surrounding neighbourhood.

3.1.3 Clustering (k-means)

The k-means algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified. It scales well to large number of samples and has

been used across a large range of applications in many different fields. These distance-based outlier detection algorithms as k-means, work on the assumption that the normal data objects have a dense neighbourhood and the outliers are far apart from their neighbours.

3.2 Filtering / noise reduction

In process industry, noise is an unwanted disturbance in an electrical signal. Noise generated by electric and electronic devices varies greatly as it is produced by several different effects. In steel industry, for example Electric Arc Furnaces (EAF) have a power supply of 10 to 100 MVA. Electrical drives of rolling mills have several huge electrical engines (e.g., a six stand rolling mill has 12 to 24 engines) with up to 5 MVA each. This leads to an electrical emission that affects the measurement devices. Finally, the measured time series contain noise, a high frequency overlay (blue) to the expected original information (red) as shown in Figure 4. There are many techniques used to filter or de-noise a time series with the aim of the removal of unwanted (not realistic) features or improve subsequent processes. A selection of such techniques is described in the following chapters.

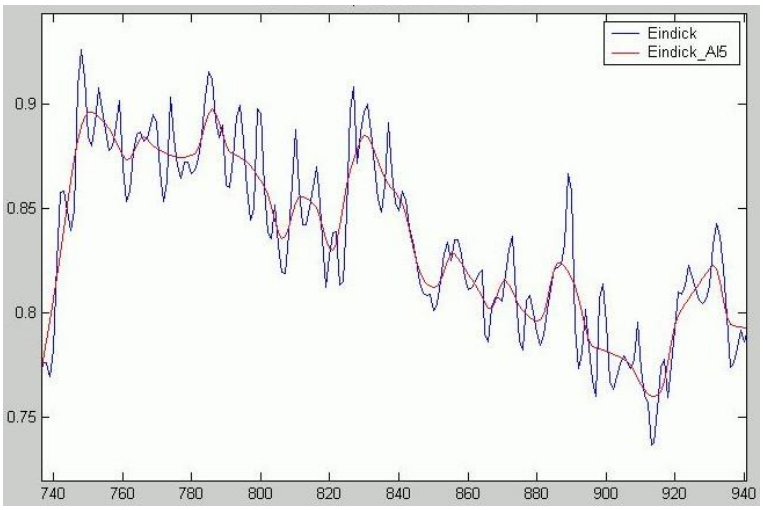

Figure 4: Time series with noise (blue) and the expected original information (red)

3.2.1 Frequency based filters

Frequency-based filters are used to isolate the frequencies that are of interest from those that are irrelevant or include faulty information. They can be used to eliminate high frequency noise or low frequency trends and keep the frequencies of interest unchanged. Typical examples are low-pass, high-pass and band-pass filters.

3.2.2 Savitzky-Golay

The Savitzky-Golay filter is based on choosing a small sub-sequence of adjacent points and approximate by a low-order polynomial the new value associated with that point. The value of

the point with the best polynomial approximation is chosen. In this work, we will pay special attention to this technique, because of its ability to maintain the overall shape of the signal.

3.2.3 Wiener filter

The Wiener filter is a linear least squares filter that can be used for prediction, estimation, signal and noise filtering, etc. It is based on statistical estimates of the signals. Having periodic signals, those frequencies that are not in agreement with the signal are eliminated, thus eliminating the noise.

3.2.4 Gaussian filter

When applying a Gaussian filter a signal is smoothed by the convolution of the original series with a kernel of adjustable length that follows a normal distribution. It is a low pass filter, in which the highest frequencies are filtered but it is applied in time domain in opposite e.g. to FFT filters.

3.3 Sub-sampling / interpolation

In industrial processes, there are many different types of measurement devices. When measuring different physical properties of a product of a process the sample rates practically used vary significantly. This can be several kilohertz for fast processes like a cold rolling mill down to 1 hertz or less for slower process like the Flash Smelting Furnace (FSF). For process analysers, the sampling time may be even longer than 10 minutes. For some data analytics or data mining methods it is necessary to have the data in equally sampled time series. This makes necessary to apply a sub-sampling (from higher to lower sample rates) or an interpolation (from lower to higher sample rates). For both cases, different methods are available to ensure neither to violate the sampling theorem nor to loose important information. A few exemplary methods are described below. Based on the time series generated by means of interpolation any sample rate can be realised by using the calculated ‘artificial’ data between the ‘real’ measurements. This is used to produce equal sampled time series even if the ‘original’ measured time series had different sample rates.

3.3.1 Linear interpolation

Isaac Newton introduced the linear interpolation that is the one of the simplest and that is probably the most commonly used in practice. Two given data points (x_0, f_0) and (x_1, f_1) are connected by a line. An example of raw data and a simple linear interpolation is shown in Figure 5.

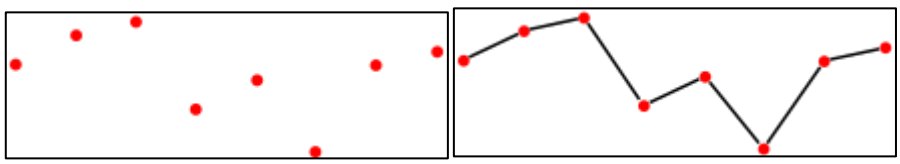

Figure 5: Time series to interpolate and linear interpolation

3.3.2 Cubic spline interpolation

Since polynomials become more and more unstable with increasing degree, i.e. oscillate strongly between the known data points, polynomials with a degree greater than five are seldom used in practice. Instead, a large data set can be interpolated piece by piece. In the case of linear interpolation, this would be a polygon course; for polynomials of degree 2 or 3, this is usually referred to as spline interpolation (Figure 6). In the case of interpolations defined in sections, the question of continuity and differentiability at the interpolation points is of great importance.

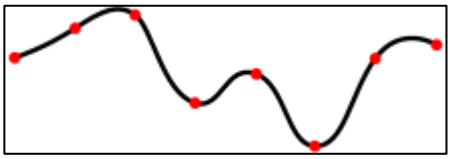

Figure 6: Cubic spline interpolation

3.4 Handling of time delays

One characteristic of process industry is the concatenation of different process steps to a process chain used to produce the final product. The following Figure 7 shows an example from steel industry (from left to right):

- Flat steel coils are welded together (1)
- processed in an electrical cleaning (2)
- through a loop accumulator (3)
- followed by a cleaning section (4)
- heating section (5)
- coating section (6)
- cooling section (7)
- temper rolling section (8)
- another loop accumulation (9)
- and finally the 'continuous' product is cut to coils (10).

In between of this chain there are often additional loop accumulators to be able to handle different process speed of the sections. They are also used to be able to handle disturbances that might appear in one section.

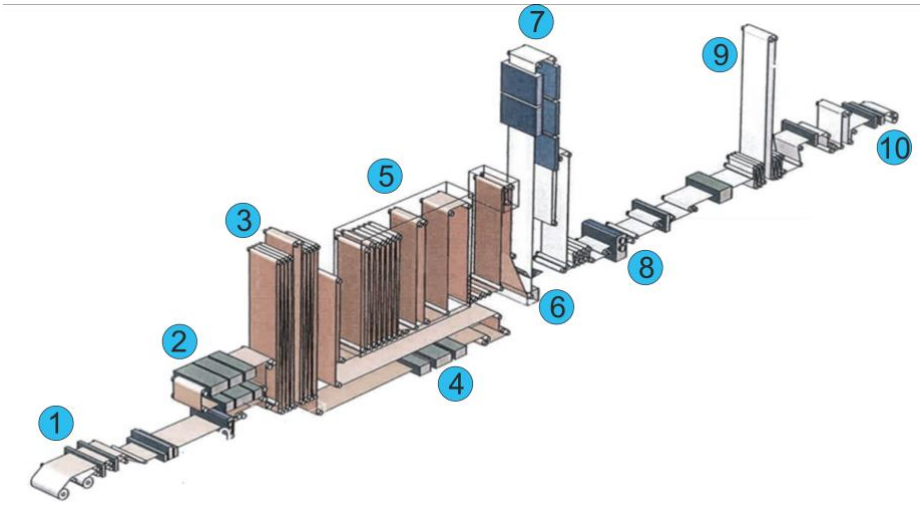

Figure 7: Process chain in flat steel production

For the analysis of data coming from such a process chain, the following problem has to be solved. Measurement devices or processing equipment providing sampled data with an equal time distance between the samples. To be able to analyse data coming from different parts of the process chain, the time stamps of the sampled data have to be adjusted so that they belong to the same position of the product. This becomes more complex, because due to the several loop accumulators and different product speed in several sections, the delay in each section is a function of the processing speed. Thus, varying time delays have to be handled to be able to provide consistent data for the data analysis task as shown in the following example.

> *A quality measurement device at position 10 (see Figure 7) that is subject of a data analysis task. For the analysis, signals of measurement devices from positions 2, 6 and 8 are necessary. The specific position of the product measured at position 10 has been measured earlier by the devices at position 8 and even earlier by the device at positions 6 and 2. To combine the right data, the values with respect to position 10 from time stamp t4 (see Figure 8), position 8 from t3, and so on, have to be combined to a vector (Figure 9). Furthermore, the delays (t2 - t1, t3 -t2, etc.) depend on the velocity of the product and the fill level of the loop accumulators in positions 3 and 9. For example, if a new coil is welded (position 1), the product speed is zero and the fill level of the loop accumulator 3 decreases. If the welding is finished, the product speed is high and the content of the loop accumulator 3 increases. In both cases, the speed in sections 4–8 is constant. This leads to the effect that for each value measured at position 10, different time delays have to be taken into consideration, when combining the data from the measurements for conducting consistent data analysis*.

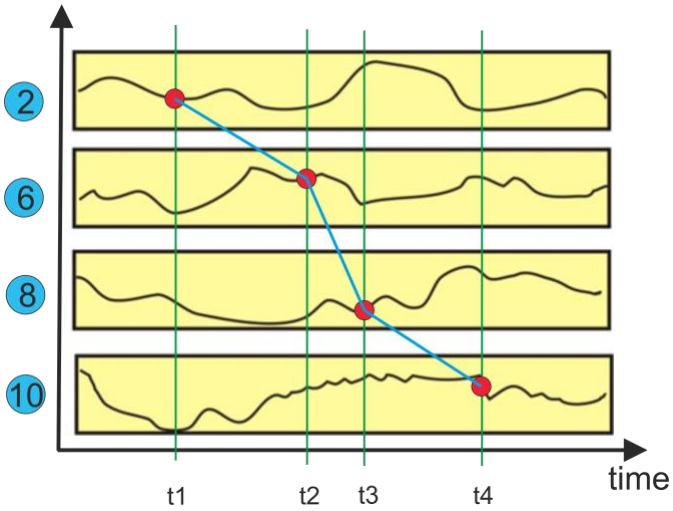

Figure 8: Example for delay handling during data pre-processing

4 Pre-processing tools from a social innovation perspective

From a social innovation perspective, a co-creation process provides relevant information for the pre-processing of data. To find out, how (future) users and other plant experts are currently handling process data and what they require from the COCOP system, TUDO and VTT conducted interviews with installation and quality managers at the steel plant of SIDENOR and process engineers at OUTOTEC in the copper case (see relevant staff described in deliverable 2.1 and 2.4).

Based on experiences with current computer systems and on expectations regarding the COCOP system, future users reported current practices of processing data. Their decisions related to process control are based on data *and* on practical knowledge. Some of the interviewed people rely more on data, some more on practical knowledge. Mostly, decisions are results of the best possible interplay of both, data *and* practical knowledge. Computer programs provide data about deviations from target state of the manufacturing process, but the installation managers need practical knowledge how to detect the causes for that.

At DSM, a COCOP partner from the chemical industry, a method of using the practical knowledge of process experts has been applied. Firstly, a data-curation procedure was applied to the raw process data. All data points that were designated as 'outliers' were discussed with the plant staff. In all cases where DSM signalled 'outliers' discussions with plant staff, the result was in fact-based conclusions about the quality of these data. An example was the fact that a sensor did not work properly in a certain month. In this way, we only removed input data when we discussed and found an objective, physical reason, why the data were not acceptable. This is not necessarily the same result as when applying purely mathematical data pre-processing tools. Subsequently, the data set was used to build a model based on artificial intelligence (AI), in this case a neural network based model. Advices for improved settings in the process are given by this AI based model by imposing certain settings, for instance a higher throughout while retaining the product

quality. This then leads to numerous mathematically generated possible solutions for process optimisation. This many solutions all fulfil the requirements resulting from the model that is purely based on available experimental data. Whereas these solutions are all correct according to the mathematical analysis, they are not necessarily correct when knowing the process operations. Also staying in a range of parameter values that have before been realized on operations gives more certainty that a solution is a proper solution for the process. The graph below illustrates this. The first optimisation revealed solutions for better operations represented by the red symbols. However, that combination of parameter values for A and X had never occurred in operations, which gave a doubt whether these mathematical solutions would also be plausible physical solutions. So all mathematical solutions generated in a next round had to (a new constraint was applied) lie in the part right of the light-blue diagonal line. This resulted in the green symbol, in the middle of the experimental data points, as a potentially feasible solution.

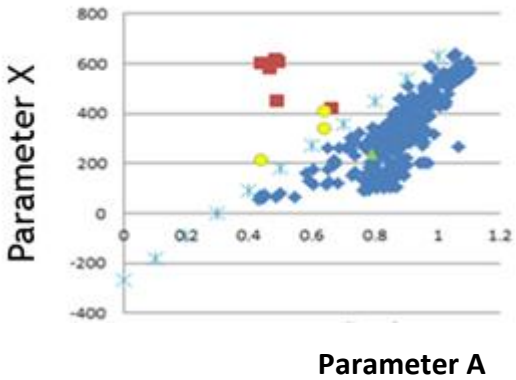

Figure 9 Staying in a range of parameter values that have before been realized on operations gives more certainty that a solution is a proper solution for the process.

In the discussions with the process experts, they gave feedback which of the possible solutions practically make sense in view of their experience. Here it is very important to differentiate between 'we think it works this way' or 'we generally know it has always been working like this' and very solid and 100% objectively correct information from the plant operations. This process involved various iterations until feasible solutions for process optimisation resulted. The engineers could not find any further arguments why these solutions could not work, and therefore this process of discussions had led to lower the barrier to the COCOP system acceptance by plant engineers.

Based on these experiences and the interview results from the steel case showing current practices of handling data, in the following Table 1 human factors requirements were derived how to (pre-) process data using the practical knowledge of the process experts.

Table 1: Human factors requirements

Requirement	Activity	Benefit
The COCOP system shall be improved with practical knowledge during the development, e.g. by excluding non-realistic solutions.	Regular meetings with developers, process experts and possibly a subset of end users will take place. They will evaluate whether the COCOP system is appropriate from a practical point of view. The developers will implement the new features that are agreed. These processes continue until an agreement between the developers and process experts is reached that the COCOP system is appropriate and helps to reach improvements of plant-wide processes.	Ensures that plant-wide optimisation brings the envisioned benefits.
The COCOP system should provide advices to improve the reliability of practical knowledge of the users.	See above	Ensure that the system fulfils the end user needs. Enhances end user expertise.

5 Conclusion

The data pre-processing is an important part in the COCOP development. Especially for data based models, (see also Deliverable D3.6) the quality of the data is crucially important for the model accuracy. Dependent of the type of data, sampling rate, measurement devices and several other influencing factors, no generic solution for the data pre-processing can be provided. For each task, a different set of methods for outlier detection, filtering / noise reduction or sub-sampling / interpolation might be necessary to reach the aims of the underlying application.

6 References

[1] See also https://en.wikipedia.org/wiki/Exploratory_data_analysis

[2] Berry, Michael J.A. and Linoff, Gordon, “Data Mining Techniques: For Marketing, Sales, and Customer Support”, John Wiley & Sons, Inc. 1997.

[3] Fayyad, Usama, Piatetsky-Shapiro, G. and Smyth, P., “From Data Mining to Knowledge Discovery: An Overview”, Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996

[4] Fayyad, Usama, Piatetsky-Shapiro, G. and Smyth, P., “Knowledge Discovery and Data Mining: Towards a Unifying Framework”, KDD-96 Proceedings, pp. 82-86

[5] Collier, K.; Carey, B.; Grusy, E.; Marjaniemi, C.; Sautter, D.; "A Perspective on Data Mining", Northern Arizona University, 1998

[6] Kadlec, P., Gabrys, B., & Strandt, S. (2009). Data-driven Soft Sensors in the process industry. Computers and Chemical Engineering, 33(4), 795–814.

[7] Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In ACM sigmod record.