Concepts for Data Evaluation
Colophon

Title
Concepts for Data Evaluation

Author(s)
J. van den Broeke (SCAN)
C. Moore (BBE)
S. Isz (Alpha-MOS)

Quality Assurance
J. van de Vossenberg (KWR)

Deliverable number
D 3.1.4

This report is:

PU = Public
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description of Workpackage 3.1, component 3.1.4</td>
<td>3</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>4</td>
</tr>
<tr>
<td>1.1 General Introduction</td>
<td>4</td>
</tr>
<tr>
<td>2 Method</td>
<td>5</td>
</tr>
<tr>
<td>3 Evaluation of Available Methods</td>
<td>6</td>
</tr>
<tr>
<td>3.1 Statistical analysis</td>
<td>6</td>
</tr>
<tr>
<td>3.1.1 PRINCIPAL COMPONENTS ANALYSIS</td>
<td>7</td>
</tr>
<tr>
<td>3.1.2 DISCRIMINANT FACTOR ANALYSIS</td>
<td>7</td>
</tr>
<tr>
<td>3.1.3 PARTIAL LEAST SQUARES</td>
<td>8</td>
</tr>
<tr>
<td>3.1.4 SOFT INDEPENDENT MODELLING OF CLASS ANALOGY</td>
<td>9</td>
</tr>
<tr>
<td>3.1.5 MULTIVARIATE STATISTICAL QUALITY CONTROL</td>
<td>9</td>
</tr>
<tr>
<td>3.2 Available Specialised Statistical Software Tools</td>
<td>10</td>
</tr>
<tr>
<td>3.2.1 Commercially Available General Statistical Software Programmes</td>
<td>11</td>
</tr>
<tr>
<td>3.2.1.1 XLSTAT</td>
<td>11</td>
</tr>
<tr>
<td>3.2.1.2 UNISTAT</td>
<td>11</td>
</tr>
<tr>
<td>3.2.1.3 Minitab</td>
<td>11</td>
</tr>
<tr>
<td>3.2.1.4 The Unscrambler</td>
<td>11</td>
</tr>
<tr>
<td>3.2.1.5 STATISTICA</td>
<td>12</td>
</tr>
<tr>
<td>3.2.1.6 SAS</td>
<td>12</td>
</tr>
<tr>
<td>3.2.1.7 SPSS</td>
<td>12</td>
</tr>
<tr>
<td>3.2.2 Evaluated Specialised Software Packages for Data Evaluation</td>
<td>12</td>
</tr>
<tr>
<td>3.2.2.1 IT SEES</td>
<td>13</td>
</tr>
<tr>
<td>3.2.2.2 CANARY</td>
<td>16</td>
</tr>
<tr>
<td>3.2.2.3 AQUARIUS</td>
<td>18</td>
</tr>
<tr>
<td>3.2.2.4 s::can vali:::tool and ana:::tool</td>
<td>19</td>
</tr>
<tr>
<td>3.2.2.5 Alpha Soft</td>
<td>22</td>
</tr>
<tr>
<td>4 Conclusions and Future Work</td>
<td>26</td>
</tr>
</tbody>
</table>
Description of Workpackage 3.1, component 3.1.4

Component 3.1.4: Concepts for data evaluation

Research activities:
Besides the monitoring of key parameters using a variety of analytical tools, and the development of additional tools, the methods for handling the data generated by analytical instruments are highly important. For example, the use of a combination of instruments that provide supplementary information can present a more complete picture on water quality than the use of a single instrument. However, in order to profit from such a battery of monitoring techniques, integrated data evaluation is required to address the large amount of data generated by such a system and to draw relevant conclusions from the data. Therefore, concepts and software for data integration and data evaluation are essential when expending the scope of applications of monitoring and control technologies. Existing concepts for data evaluation will be identified and the most suitable one(s) will be evaluated.
1 Introduction

1.1 General Introduction

Analytical systems and sensors can be used to generate large amounts of information. These are typically stored either as single files or are compiled in a database. Typically, this information is used either to monitor processes (e.g. using temperature, pH, conductivity, turbidity) or to extract key parameters using highly skilled knowledge (e.g. when using analytical systems in the laboratory). In some applications or for some quick assessments, this information needs only to be compared and classified. However, much of the data is only briefly looked at, and if nothing unusual is seen at first glance it is stored forever. This is the results of the lack of appropriate, user friendly tools for data evaluation and results in a ‘data graveyard’.

By using advanced, (partially) automated, statistical software it is possible to extract much more information from the data than possible when looking at single parameters. Pattern recognition algorithms within a software allow quick comparison and identification of samples or provide information on changes in the water quality in source waters, distribution network or provide information on treatment performance. Typically, these advanced programmes use some form of multivariate statistics to analyse the data and separate the most important components.

These statistical methods are offered in various forms. The most basic ones have to be configured manually in statistical / mathematical software packages such as Mathworks Matlab or Statistica. These are powerful tools, but mainly suited for academic use. Partially and fully automated programmes for data evaluation are offered as well. Typically, these are provided by manufacturers of measurement equipment, and they are specialised in processing the data supplied by the equipment of this particular manufacturer. Therefore, their general use is limited. Finally, some software tools exist that transcend this manufacturer dependence, and which are open enough to allow evaluation of any data, as long as they are presented in a compatible format. The document focuses on the discussion of a few basic principles in data evaluation as well as an assessment of a number of software packages of the latter category.

It must be noted, that any evaluation of such powerful software tools can only be done based on real experience in working with the individual softwares. Information taken from brochures or marketing materials have very little value. Therefore, this evaluation is restricted to the packages that the WA3 partners had experience with at the time of writing. It is realised that there are additional software applications available on the market, but due to the lack of experience with these within the project team, it was not possible to present an objective evaluation of these software packages here.
2 Method

All evaluations of the monitoring technologies have been made by the participants in WA3. Although there is a large expertise among the WA3 partners, it should be kept in mind that these evaluations are based on personal experience and judgement and should not be taken as absolute. Evaluations focussed on the practicality of the methods / products evaluated. Many more techniques have been published, especially in academia, but these often remain inaccessible for use in day to day operations of drinking water works. Therefore, only methodologies that have found their way into commercially available software packages are discussed.

The basis for this report is a table that was based on the table for the TECHNEAU report Monitoring and control of drinking water quality - Selection of key parameters (Deliverable 3.1.3; Mons et al., 2007, see www.techneau.org ). The table was adapted to suit assessments of software packages. To prepare evaluations in a uniform format, and make sure that evaluations from different partners include information on similar aspects, a standard evaluation form was prepared. The basic evaluation form with the table is shown in Fig 1.

Fig. 1 Evaluation form
3 Evaluation of Available Methods

3.1 Statistical analysis

This section of the report describes several algorithms and methods that are applied in data evaluation. It explains the basics of multivariate approaches.

Comparing patterns of data produced by analysers or sensors can be done visually by comparing graphs and charts depicting variable (sensor) response for various samples. Comparing two variables (sensors) responses charts is relatively easy. Differences on each sensor may be identified and quantified. If differences between sensor responses are observed, the user has to determine which differences are important, and to what degree.

In order to obtain DISCRIMINATION (for two samples or water qualities to be called different), we need to understand how large the individual sensor response differences need to be, and how many sensors need to exhibit differences. For this step it is desirable to use more complex data analysis techniques to compare data sets as a whole. This implies reduction in the number of variables. As an example, data reduction allows responses of a high number of variables (sensors for example) to be processed and displayed in a two or three dimensions map.

Unsupervised learning techniques like Principal Component Analysis (PCA) are widely used to explore data and to assess discrimination performances.

With trained or supervised techniques, mathematical models are developed to be then able to identify unknown samples into one of the training group. IDENTIFICATION may be obtained by using:

- Discriminant factor Analysis;
- Partial Least Squares;
- Soft Independent Modelling by Class Analogy.

Statistical methods are useful for extracting and interpreting the pertinent information contained in a data base. Before performing data processing, the user needs to create a data base that is generally presented as a matrix $X$ of $n$ rows and $p$ columns. The generic term $X(i,j)$ of $X$ is the value taken by the $j^{th}$ variable on the $i^{th}$ individual or object. In the FKP Software, the data base $X$ is constructed using different modules of « LIBRARY » menu. The variables are the sensors and the objects are the analyzed samples.

In the next sections, a description of the methods is performed before guiding the user onto the software.
3.1.1 **PRINCIPAL COMPONENTS ANALYSIS**

Principal Components Analysis (PCA) allows data exploration and was proposed in 1933 by Harold Hotelling. The main ways to make the exploration is:

- **An assessment of the similarities between samples:**
  Which samples are different or similar? Are they separated in homogeneous groups?

- **An assessment of the relation between variables (sensors or variables).**
  Which variables are linked or opposite? Are there extremely correlated variable groups?

Another aspect of the second objective is to summarize the set of variables by a small number of synthetic variables, the “Principal components” which are linear combinations of the original variables.

PCA determines subspaces that summarize information contained in the initial database. The two main objectives are:

- Reducing number of variables;
- Eliminating redundancy.

Presentation of the results of the analysis is done in a two or three dimensions space relatively to the chosen components (from $C_1$ to $C_n$). Components are classified depending on the level of information they explain.

3.1.2 **DISCRIMINANT FACTOR ANALYSIS**

PCA allows the extraction of information in a data table. In particular it summarises the similarities between individuals but also the relations between variables. In some situations, however, additional information is available on the individuals. In particular, we may have knowledge about an « a priori » partition of the individuals set and know the group of each individual. In this case, each group represents a territory and each individual is a sample. In this situation, we try to find « descriptive » variables predicting this partition.

The algorithm consists of looking for new variables, linear combination of the descriptive variables, which separate as good as possible the different groups of the partition. That is the first « descriptive » objective of the Discriminant factor Analysis (DFA).

Second objective of this method is « decisional » and allow group prediction of a new individual.
The set of points $N_x$ is partitioned on $q$ subsets $\{N_x^1, N_x^2, \ldots, N_x^k, \ldots, N_x^{q-1}, N_x^q\}$ that represent different qualities, origin descriptors. The objective of DFA is to look for new variables $\{F^1, F^2, \ldots, F^k, \ldots, F^q\}$ corresponding to directions which separate groups.

The second objective of DFA is to solve the classification problem: How to allocate unknown samples to defined groups? For each unknown sample the distance to the gravity centre of each group is computed. The sample is then identified as part of the group associated to the minimum distance. A validation of the model allows the qualification of model predictions. Two validation methods are generally used:

1. **Leave one method**: Each sample is removed from the data set and considered as unknown. A DFA is performed with the other samples and the removed sample is identified in one of the groups. Total score of all samples is computed. Generally, a score of 95% is required to validate the model.

2. **Bootstrap method**: Half of the samples are randomly removed from the data set and identified in one of the predefined groups. This operation is repeated until convergence. The average of the obtained scores is calculated. It gives an idea about model robustness. Generally, a score of 90% is required to validate the model.

### 3.1.3 PARTIAL LEAST SQUARES

This method is used to extract quantitative information. In volatile organic compounds (VOC) analysis by chemical measurement and by panel evaluation, e.g. there are mainly two types of quantitative information:

- Concentration of specific compounds (e.g. concentration of benzene in water…);
- Sensory panel score.

The objective is to build a model that is able to predict this quantitative information for each sample to be analysed.

The PLS algorithm is based on linear regression techniques. $Y$ is the matrix containing quantitative measurement and $X$ is the matrix built with detector measurement (gas sensor array, mass spectrometer, spectrophotometer, multiple single sensors). As a simplified explanation of PLS model, it can be described as looking for a $B$ matrix that minimises distances between $Y$ and $Y'$ with $Y'=X.B$.

After building the model, the $B$ matrix is used to predict quantitative information contained in an unknown sample. The measurement matrix is then multiply by $B$ to obtain the prediction. Some values are also of interest to assess performances and accuracy of the model (correlation coefficient…).
3.1.4 **SOFT INDEPENDENT MODELLING OF CLASS ANALOGY**

The interest for this method comes from the ability to build an optimized model by taking into account its ability to identify only one group: good or bad products for example.

The model is going to identify unknown samples as belonging or not to one and only one group. The mathematical process is comparable to a combination of DFA and PCA. The basic explanation is provided hereafter.

The model learning database is relative to one group. After running PCA on the data, the objective is to find the best synthetic subspace which best contains samples. Each sample is defined by two values:

- Its projection on the subspace S: y1.
- Its projection on the S* orthogonal subspace: y2.

As a consequence, set of points in the matrix X (detector measurement) induce two new ones:

- One in S: Nx(S);
- One in S* called residual set: Nx(S*).

We can now define identification criteria. For each unknown sample y, two values are computed:

- The norm of the residual part y2;
- The Mahalanobis distance of y1 part to the gravity centre of Nx(S).

Two theoretical values are defined as threshold for the identification.

3.1.5 **MULTIVARIATE STATISTICAL QUALITY CONTROL**

The MSQC is used most often as several variables are needed to assess a product. MSQC is an extension of the univariate Statistical Quality Control (SQC) usually used when only one variable or measure is taken into account.

- Statistical Quality control (SQC) - one variable

One method used for many years to detect significant deviations from acceptable production performance is Statistical Quality Control (SQC). The traditional approach charts single variables and compares trends with historical behaviour to judge deviation from the norm. Because each variable is considered separately, this form of SQC is called univariate. Traditional univariate SQC is suitable for selected batch system problems. However, such processes in general pose challenges that make univariate SQC inappropriate.

For example, there can be a plethora of variables to record. This necessitates the tedious and time-consuming interpretation of multiple charts. Other problems include the fact that you can never achieve steady state in batch operation. Further and importantly, deviations in process performance are
often the result of interactions among variables. Charting a single parameter and using univariate SQC cannot identify these variations.

To overcome the problems associated with univariate SQC, multivariate SQC (MSPC) techniques have been developed and subsequently applied to batch systems. Multivariate SPC takes into account the interactions among variables and significantly reduces the number of charts the process operator must observe.

- Multivariate SQC (MS QC) with several variables or How to condense the variables

The most straightforward approach is to use principal component analysis (PCA). The PCA approach compresses the dimension of the process data and through transformation develops new "variables" that are essentially weighted combinations of the original process variables. These new variables, principal components, describe decreasing amounts of variation in the process subject. If correlation exists in the data, you can represent most of the process information with fewer principal components than process variables.

This data compression leads to more effective process data charting. A few multivariate charts' displaying relationships in principal components replace many univariate graphics. Deviations in these charts indicate process divergence from normal behaviour. In addition, you can reconstruct the data from the principal components, with large reconstruction errors again being an indication of process deviation.

3.2 Available Specialised Statistical Software Tools

Statistical evaluation can be performed using various software tools. The different packages available can be grouped into three classes:

- mathematical software programmes, which are completely open. These can be used to perform any calculation required, but need a highly skilled user as everything has to be written by the user. Examples: Matlab, Statistica,...
- statistical software programmes: these programmes are specialised tools for a limited number of statistical evaluations. They offer a reduced functionality but contain pre-programmed algorithms. Therefore, user skill required is lower.
- specialised software packages: these are software packages that are focused on evaluation of data produced by analytical instruments and/or sensors. Their functionalities are reduced further, but the functionalities included are focused on extraction information from the type of data known to be produced by the analytical equipment. Also, these software tools have a much simpler (graphical) user interface, which makes them accessible to non-academic users as well.
3.2.1 Commercially Available General Statistical Software Programmes

This section presents an overview of commercially available General Statistical Software packages. This overview is based on publicly available materials. The descriptions have been taken from the software distributors, some of the information was found on review websites and Wikipedia. The descriptions are not based on personal experience of the authors. This overview, although extensive, does not aim to be complete but only presents a number of software packages as example of what is available on the market today. See [http://en.wikipedia.org/wiki/Comparison_of_statistical_packages](http://en.wikipedia.org/wiki/Comparison_of_statistical_packages) for a broader overview of features in many more statistical packages than described here.

3.2.1.1 XLSTAT

[http://www.xlstat.com](http://www.xlstat.com)

XLSTAT is a data analysis and statistical tool for Microsoft Excel. The XLSTAT add-in offers functions to enhance the analytical capabilities of Excel.

3.2.1.2 UNISTAT


Unistat is a statistical package that also can work as a Microsoft Excel add-in. It contains up to 136 procedures, depending on the version. It has 2D and 3D graphical output. It is smaller than e.g. Statistica.

3.2.1.3 Minitab

[http://minitab.com](http://minitab.com)

Minitab Statistical Software is used for teaching statistics and other quality improvement projects. The package contains data management, data manipulation, and spreadsheet-like data windows.

3.2.1.4 The Unscrambler

[http://www.camo.com/rt/Products/Unscrambler/unscrambler.html](http://www.camo.com/rt/Products/Unscrambler/unscrambler.html)

The Unscrambler is (free-to-try) software for Multivariate Data Analysis (MVA) and Experimental Design. The software is designed to find patterns and relationships in data that cannot be found so easily by eye or with traditional statistical methods. The software is predominantly used in spectroscopic applications.
3.2.1.5 STATISTICA

http://www.statsoft.com

Statsoft Statistica is a large general statistics package like SAS and SPSS, and it has more options than most other statistical packages described here. Like the other large packages, Statistica provides data analysis, data management, data visualization and data mining procedures. Its techniques include predictive modelling, clustering, classification and exploratory techniques.

3.2.1.6 SAS

http://www.sas.com/technologies/analytics/statistics/stat/

SAS/STAT is the statistics part of the larger SAS package that has been on the market for over 30 years already. It has just as many functions as Statistica and SPSS, although it has been criticized for its graphics and lack of debugging tools.

3.2.1.7 SPSS

http://www.spss.com/

The first release of this package was in 1968 and designed to run on mainframes. It still is one of the most widely used statistical programs in social sciences. Nowadays it can also run on a PC under different operating systems. Like SAS and Statistica, SPSS is a large general statistics package with many functions.

3.2.2 Evaluated Specialised Software Packages for Data Evaluation

Below are presented the datasheets for a number of commercially available software packages that can be used for evaluation of data from single and multiple sources. All these software packages have a principle functionality for the detection and/or identification of changes in the composition of the medium being monitored. This is done by integration of multidimensional data, provided either by multiple instruments or by multiple inputs from a single instrument. Most of these software packages have been developed by manufacturers of sensors and analytical equipment. They can, however, be used independently of that manufacturers’ equipment.

The reason for manufacturers to develop their own software packages must be noted here as well. Although, as described in section 3.2.1, many commercially available applications exist, these are not well suited for use in combination with sensor systems. The reasons for this are the following:
a) the available software packages are not focused on the features needed for a particular analytical system. As a result they offer too many features, resulting in poor usability, or they lack specific features that are essential.

b) the packages available are not suited for use on the controllers typically delivered with analytical systems. They are mostly Microsoft Windows based interfaces that use small icons and require keyboard inputs. Also, they are over complicated for the task at hand, therefore making operation by non-specialist personnel difficult.

c) software license costs will increase product costs significantly. Even at relatively low numbers of units sold, writing own software applications is more economical. Add to this the points a) and b) explains why practically all advanced equipment come with their own software.

3.2.2.1 IT SEES

Prepared by: bbe Moldaenke GmbH, Germany

Specifications:
IT SEES runs on the Windows platform and can be integrated into other operating systems.

Input data sources:
Readings from various (existing) sensors e.g. temperature, absorption, pH redox, etc via a SCADA system; CSV, TXT file formats for database import is implemented. Drivers can be made available for different database systems.

<table>
<thead>
<tr>
<th>Text File Format example</th>
</tr>
</thead>
<tbody>
<tr>
<td>timestamp</td>
</tr>
<tr>
<td>dd.mm.yyyy</td>
</tr>
<tr>
<td>hh:mm:ss</td>
</tr>
</tbody>
</table>

Scheme 1: Schematic diagram of set-up.
Description:
The software was developed within the EASE Project (Federal EPA Berlin, EPA Hamburg, bbe Moldaenke GmbH). It is commercially available from bbe Moldaenke GmbH. IT SEES is a tool which can analyse all kind of data for sudden changes, find correlations to changes in various parameters and estimate whether an alarm situation is at hand. It is to help users of different measuring instruments to assess the ongoing process in their observed water systems. The described algorithms are made to detect changes in each observed signal. The user can enter lower and upper limits in the evaluation, statistical short-term evaluation as a standard deviation using algorithms with and without drift elimination, and long-term evaluation. Usually events from different signals at the same time period are necessary to indicate a reportable alarm. A weighting system is applied to the events shown by each signal.

![Figure 1: Upper left: Alarm Index – results of evaluation of all sensors – each measurement parameter adds a certain number of toxicity points to the index.](image)

Evaluation:
The software is a unique and relatively simple detection tool for multiple sensors. It is in use in environmental measuring stations. It can be used by non-skilled and semi-skilled personnel since data input is pre-defined and very few parameters need to be configured by the user.
<table>
<thead>
<tr>
<th>Criteria</th>
<th>1 = very bad</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5 = very good</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technical specifications</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X robust, stable Windows programme with little user input or configuration</td>
</tr>
<tr>
<td>Robustness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Operational specifications</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X Windows-based software with graphic interface</td>
</tr>
<tr>
<td>Ease of use</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5 = very good</td>
<td></td>
</tr>
<tr>
<td>Interoperability with external data sources</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X database information can be imported from other databases</td>
</tr>
<tr>
<td><strong>Costs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X license costs</td>
</tr>
<tr>
<td>Hardware</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>runs on a Windows PC</td>
</tr>
<tr>
<td>Operational costs</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>none (occasional updates)</td>
</tr>
<tr>
<td>Recommendation for use in SSS</td>
<td>YES or NO:</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td>Water works, industrial plants, environmental monitoring</td>
</tr>
<tr>
<td><strong>Overall conclusion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>a continuous, real-time, online and offline tool for environmental monitoring stations or industrial plants to evaluate data of multiple sensors, with the highest sensitivity and the lowest false alarm rates possible</td>
</tr>
</tbody>
</table>
Prepared by: s:can Messtechnik

Website: http://www.epa.gov/NHSRC/news/news122007.html
http://www.epa.gov/oamcinc1/0810286/canary1.pdf

Purpose and use: Event detection system

Technical specifications:
CANARY is a Windows program that runs on nearly any computer. It is written in the MATLAB® m-code language but also employs some Java programming language. The typical user will not need to know MATLAB®; however, modifications to the source code require knowledge of MATLAB® and/or Java. CANARY is available via the Lesser GNU Public License for research purposes; it is not intended to replace any commercial software, but rather to serve as an aid in developing software.

Input data sources: Sensor readings (single parameters and/or spectra) from SCADA system or offline via CSV files.

Description:
The software and its source code are available on the US EPA website. Canary was developed by the Sandia National Laboratories for the EPA. It is designed to work with SCADA systems as well as with offline data. It is currently using different techniques to determine whether a measurement represents a baseline measurement or an anomalous situation. These predictions are performed using linear filters or by repeating the last measurement and then comparing it with the actual measurement and nearest neighbor pattern recognition. The alarm probability is then evaluated by using a binomial event discriminator, which calculates the probability of a certain number of unusual measurements (outside the tolerances) within a certain time window. Canary is an open system that allows external researchers or developers to plug in their own algorithms and to test them in the Canary environment.

Evaluation:
The software is very flexible and powerful. As the source code is available, it can be adapted to individual needs. It is not user friendly and more a scientific tool than user software.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>1 = very bad</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5 = very good</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robustness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>detection algorithms OK, but no filtering of input data</td>
</tr>
</tbody>
</table>

Concepts for Data Evaluation  
© TECHNEAU  - 16-  July 9, 2009
occurs, all outliers and noise in included and hamper event detection.

<table>
<thead>
<tr>
<th>Operational specifications</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ease of use</td>
<td>X</td>
<td></td>
<td>no user interface available</td>
</tr>
<tr>
<td>Interoperability with external data sources</td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Costs</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Software</td>
<td>X</td>
<td></td>
<td>software is free.</td>
</tr>
<tr>
<td>Hardware</td>
<td>X</td>
<td></td>
<td>runs on a normal PC</td>
</tr>
<tr>
<td>Operational costs</td>
<td>X</td>
<td></td>
<td>no operational costs</td>
</tr>
</tbody>
</table>

| Recommendation for use in SSS | YES or NO: | No |

<table>
<thead>
<tr>
<th>Overall conclusion</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Canary is a powerful scientific tool, but in its current development state it is unsuitable for use by non-scientific users.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.2.2.3 AQUARIUS

Prepared by: s::can Messtechnik

Website: http://www.aquaticinformatics.com

Purpose and use: management and analysis of water quality, hydrology, and climate time-series data.

Technical specifications: AQUARIUS runs as a stand-alone application on MS Windows machines or on Apple Macs running Bootcamp. AQUARIUS
3.2.2.4  s::can validation and analysis tool

Prepared by: s::can Messtechnik

Website: www.s-can.at

Purpose and use: Event detection system

Technical specifications: PC, industrial PC or s::can terminal. Interface can be any device with browser and network connection.

Input data sources: Direct Sensor inputs (analogue signals, MODBUS), Database, ASCI files

Description:
The s::can EDS works with a two step approach. A data validation module is used as pretreatment of the input data to minimize the influence of bad measurements on the EDS, before the EDS is analyzing the input data with a multitude of data analysis tools.

In order to optimize the performance of an EDS, the input measurement data needs to be as clean as possible. This means firstly, that all sensors need to be installed and maintained as good as possible and secondly, that bad measurements that could not be avoided are filtered out. Both can be achieved by a data validation tool. A data validation tool will mark measurements that are suspected to be bad measurements as not reliable. These measurements will be excluded from the input of the EDS. If there are frequent bad measurements from a sensor, the operator will be notified by the validation tool early, which will help to keep the times of non optimal operation of sensors short. The data validation is done by self-adapting statistical models and by time series models.

The clean data stream is then analyzed by the s::can EDS using a number of simple and robust tools that are each made for rather simple tasks, each extracting one special feature from the input data. The information, whether specific features are present or not, is then condensed to an overall event probability. The EDS tools are based on self-adapting statistical models, time series models and nearest neighbor pattern recognition.

Costs
The standard software package for sensor managment and data acquisition will be supplied with s::can instruments or can be acquired separately. License costs are expected to be in the order of 2000 Euro. The data validation and alarm generation modules are additional packages with their own costs. Expected price levels for these packages will be around 1000 euros per package. Additionally, a service contract for the software, which ensures the software will be updated with new features as they become available will also be offered. Prices for stand alone packages, i.e. for use in a central database, will be quoted to customers individually.
Figure 1: Pattern recognition and allocation of water types / alarms levels to different water compositions.
Figure 2: Step detection.

**Evaluation:**
The software is being designed to be easily used by low qualified personnel. It is very robust and needs no configuration by the user. Its weakness is therefore that it is not very flexible and not useful as a scientific tool.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>1 = very bad</th>
<th>2</th>
<th>3</th>
<th>4 = very good</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technical specifications</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Robustness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td><strong>Operational specifications</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ease of use</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X designed for ease of use</td>
</tr>
<tr>
<td>Interoperability with external data sources</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>designed to work with online data and not optimised for analysis of historical data, therefore only a limited number of data formats are supported</td>
</tr>
<tr>
<td><strong>Costs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X License costs</td>
</tr>
<tr>
<td>Hardware</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X runs on a normal PC</td>
</tr>
<tr>
<td>Operational costs</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X license maintenance cost to keep software up to date.</td>
</tr>
<tr>
<td><strong>Recommendation for use in SSS</strong></td>
<td>YES or NO:</td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Overall conclusion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>vali::tool and ana::tool are useful tools optimised for operation on monitoring stations for real time data validation and alarm generation. The tools are robust and designed for highest sensitivity with low false alarm rates. It is meant as an operational tool, and therefore its use for scientific purposes is low.</td>
</tr>
</tbody>
</table>
3.2.2.5  Alpha Soft

Prepared by: Alpha MOS

Website: www.alpha-mos.com

Purpose and use:
Control and pilot gas and liquid analyzer (electronic nose and tongue) based on gas sensors, liquid sensors, MS and GC.
Analyze obtained results with statistical neural network to compare samples:
- Samples differenciation
- Samples quantification
- Quality and process control
- Sample ageing
- .....  
Can be either on line data treatment or after analysis

Technical specifications:

system requirements:
- operating system: Windows XP or Vista
- interface: standalone program
  - HMI (Human Machine Interface) in Visual Basic or Visual C++
  - Access to the data via Visual Basic
  - Control of the different subsystem drivers and statistics via C++

- database types: MS Access databases
- data format: XML files
- extension to other software: export in .txt format / AIA format .cdf

price: 5000 €

Input data sources:
- Sensor readings: gas sensors, liquid sensors, MS and GC.
- Measurements of resistance, potential, .....  

Description:
The system is controlled under Window XP / Vista / 2000. The software is developed within C++. Several users can be declared with a given profile defined by the software administrator. Routine user will use the instrument as a QC tool, whereas the developer user will have access to all tools available. Including the administrator, three profiles are therefore available. The developed software allows to control the different subunits and to analyze the collected data. This software can be either dedicated to the RD or the QC departments as corresponding statistical tools are available. The developed methods are using the UML (United Modelling Language) formalism. The ER (Relation Entity) formalism is used to build a data base to
generate related databases. XML (Extended Metafile Language) is used to save the generated data.

The software architecture can be described as follows:

Data processing includes:
- Raw data storage in specific files
- Creation and storage of a library which corresponds to a table of the different analyzed samples with the different technologies. For each sample, for each sensor, a key result is provided which corresponds to the equilibrium stage, pick height, ....
- Different statistical tools:
  - PCA (principal Component analysis)
    It allows to explore the collected data by determining subspaces that summarize the information in the initial database. Usual representations are 2D or 3D maps
  - DFA (Discriminant factor Analysis)
    The system is first trained with defined and fully characterized samples set. Then DFA allows identification of unknown samples.
  - SQC (Statistical Quality Control)
These techniques have been developed and subsequently applied to batch systems. They take into account the interactions among variables and significantly reduce the number of charts the process operator must observe.

- **PLS (Partial Least Squares)**

  This method is used to correlate the system measurement with quantitative sample characteristics such as concentration, sensory evaluation. The objective is to have a model able to predict quantitative information of each analyzed samples.

- **Shelf life**

  This method allows to follow product stability over time

**Status of the technique**: *commercially available*

In use worldwide

In native languages: English, Spanish, German, Japanese, Chinese, French
new language can be added upon request.

Reference to vendors
Alpha Soft

standards/norms
archive, traceability, data integrity, reports,
21 CFR part 11 activate or deactivate mode
GLP

Evaluation:

This software is a user friendly allowing one to have access to complex data treatment without high skills in statistical data treatment.

On line and at line sample evaluation can be performed.

Software: Alpha Soft

<table>
<thead>
<tr>
<th>Criteria</th>
<th>1 = very bad</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5 = very good</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technical specifications</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Robustness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Operational specifications</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ease of use</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interoperability with external data sources</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Costs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>Cheaper than others on market</td>
</tr>
<tr>
<td>Hardware</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Operational costs</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Recommendation for use in SSS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YES or NO:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Overall conclusion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friendly use software to collect data and analyze them with statistical data treatments to provide a clear information on samples: types, concentration, sensory qualification or intensity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

more information on request
4 Conclusions and Future Work

The objective of the activity described here was to create an overview of methods and tools for evaluation of complex datasets. Combining literature information and personal experience from the TECHNEAU WA3 partners such an overview was created. It was found that many different approaches, especially to statistical data evaluation, are available. It was found that most statistical software packages are very powerful, but only suited for use by highly skilled personnel. More user friendly software tools are offered by instrument manufacturers. These tools are often focussed on the instruments provided by the manufacturer, but they also offer interfaces to databases or text files. Thus they can be used for data analysis, detection of changes in water quality, classification of water types etc. Also, such systems are generally provided with a graphical user interface which makes them suited for use by waterworks personnel.