

General technical terms and conditions for NeuroModel studies

Models and model validity

At the beginning of the modelling work, client and atlan-tec Systems GmbH (hereinafter called ats) agree upon which data of which periods of time shall be used for modeling and for acceptance and verification/validation of the models.

The available data from these periods must be complete and contain all online tags of the DCS, and all laboratory data, which are relevant to the problem. In addition, the client will inform ats on request about all operational peculiarities during these periods, such as in particular shutdowns, raw material changes, product changes and other changes in the process, which are not clearly identified in the records.

Client and ats agree that the modelling data does not contain all possible operating conditions of the process. Thus, the modeling result does not necessarily represent all possible operating states of a process. Models will only be created for those value ranges and thus operating conditions which are represented in data records in sufficient numbers and quality. Which ranges and conditions these are, depends on the observed frequency of values for the individual variables in the underlying records. In general, all records are removed in which at least one variable is in a range which is given in less than 90% of the usual operating conditions. This naturally limits the informational value and the validity of the models.

Since the characteristic of the manufacturing process of various products differs on individual production lines, it is possible that this study will recommend the separation of models by products; i.e. that for each product its own model must be created. This separation in partial models improves the accuracy of neural models considerably and therefore must be ensured (be performed) for future practical applicability of the system. The detection of the operating state of the system or a characteristic phase of a batch and thus the currently manufactured product is realized by the identification of different ranges of values.

Consequently, this study may deliver the necessary models for representative products or ranges only instead of for all possible products. A feasibility study such as this will usually provide models for products and ranges which are represented in the data in sufficient number and quality. Which value ranges these are is decided by the project engineer who will consider the observed frequency of values of individual variables in the data provided. In general, products which make up at least 20% of the production time will be examined. This will narrow down the informational value and the validity of the models, usually resulting in a model which represents only the same or similar operating conditions of the process which were the basis of the training.

The extension of the models is carried out during the (possibly later occurring) online-implementation of a model in which, as a rule, the online training module *NeuroModel® script online* is used to update the models on a regular basis with new data and operating ranges.

Neural modeling, like any statistical method for modeling, is not a deterministic process that takes into account physical conditions. Internally, the model is based on unitless variables without any physical significance, which only need to be reproducible. Thus, the replacement of a sensor or an analyzer that changes the range of values, or the change of a process step has a serious impact on the accuracy and validation of the model. As a rule, the model must be adapted by introducing scaling factors or rebuilt after a change in procedures, variables or sensors on which the model is based. Therefore, if the study or a project step is performed with databases based on a certain configuration of the process and the sensors, and then, after completion of the work (modeling) changes are made to these parameters, the model will under certain circumstances be rendered invalid. Thus, it may be necessary to redo a part of the work subject to a fee.

If there is a change in the ranges of values of individual variables, the measurement method or a process change, these changes must be reported to ats at the time of implementation so that no work resulting in unusable models is carried out. Should the client not provide this report, increased costs apply. If process changes occur in this manner after completion of a project phase, adaptation or repetition of this phase will result in additional costs for the client.

Statements which are derived from models are correct only in the vicinity of actual operating ranges. The farther the model moves away from operation ranges actually considered in the modeling, the more uncertain will the model statements become with respect to sensitivity and causality.

If the number and the density of valid data sets used in the modeling is smaller than statistically necessary, high quality of the models cannot be ensured. This is not a fault of this specific modelling method, but of the data records made available.

The warranty for models which have been successfully created by ats and accepted by the client becomes invalid as soon as a change of process, value ranges of certain variables, principles of measurement or position of transducers occurs, as well as at the moment of failure to report such a change by the client.

Acceptances always refer only to periods before a change of process, value ranges of certain variables, principles of measurement or position of transducers. Changes of such parameters after the execution of the work by ats which seem to make an acceptance test with new records impossible, these records may not be used for acceptance tests.

In this case, only data sets from periods before the change of process are relevant.

Required process data

For the modeling of the process with neural process models, simultaneous and equidistant data records with time stamps from the process are required, as well as laboratory analyzes. Alternatively, the client may provide data sets with predefined batch fingerprints.

If laboratory samples are taken, providing intermediate or output variables, it will be necessary to precisely note the time of sampling. If inaccuracies are to be expected, they should be quantified as accurately as possible.

For this reason, it is recommended to be flexible about sampling schedules. Instead of setting samples for specific times, provide a time frame for the sampling; however, with the statement that the time of sampling must be recorded accurately. By this measure, a higher accuracy of the listed time can be achieved, with less frequent incorrect notations.

Data formats

To carry out a study with neural networks, process data and laboratory or analyzer data are needed. These data are provided to ats by the client. The following basic requirements apply to data provided by the client and transferred to ats:

Deviations of the data sets from these specifications are relevant for cost increases, which are billed after the data review by ats to the client, because the reformatting is considerably time consuming and expensive. These basic requirements are completely independent of the data format provided:

- The columns (fields) include exactly one variable or time/date each
- The rows (data sets) each contain one value for each variable or timestamp field, always correctly assigned to their respective columns or fields. (Data gaps are allowed, showing a NULL value, but double occupations of fields are not allowed.)
- Data lines must be preceded by one line with column names
- The data lines may be preceded with the physical units of the column a row.
- Column names must not be assigned twice. Each column name and thus variable name must be unique in the entire database
- The date has the following format (standard in Germany): "dd.mm.yyyy hh:mm:ss" (seconds and leading 0 for hour, day and month are optional; the date must be four digits, and the hour is on the 24 hour scale)
- It is **not** permitted to use numerical values to mark empty cells (such as -99 999 or even 0). Empty cells remain empty. If the measured value is actually 0, of course is a 0 is entered.
- All data, including all process data and all laboratory data, regardless of file format, must be given in a single file (and a single table within this file)
- The laboratory data must be entered in a separate column, but in the same table in line with all other measurement data. If the laboratory data time stamps do not match the time stamps of the other measured values, they should be in the line with the closest time stamp.
- The time intervals between the lines must be equidistant, i. e. the acquisition of measurement values must be performed at regular intervals and the data must be recorded accordingly.

Time and costs incurred due to faulty media or files can cause cost increases.

If laboratory samples are taken, providing intermediate or output variables, it will be necessary to precisely note the time of sampling.

For this reason, it is recommended to be flexible about sampling schedules. Instead of setting samples for specific times, provide a time frame for the sampling; however, with the statement that the time of sampling must be recorded accurately. By this measure, a higher accuracy of the listed time can be achieved, with less frequent incorrect notations.

The use of **data acquisition systems** with so-called "compression algorithms", such as the "PIMS" systems of Aspen (infoplus 21) or OSI (PI system), may result in data which is unusable for any systematic analysis. This can make modeling impossible. If data from these systems are used, their "compression algorithms" must be switched off **completely**.

For submitted data, following requirements apply:

All files and tables must be precisely described in an accompanying document (textfile or the same format as the data), documenting content and structure in detail.

The transferred records must have one of the following formats, unless a written agreement specifies atlan-tec's responsibility for converting the data files:

Access format (only accepted are: Access2007 till Access2010)

- The column type can be only "date" or "double". In exceptional cases, for pure texts, the type "text" may be used!

Excel formats (only accepted are: Excel 2007 till Excel 2010)

- The data cells must all be formatted as "number"
- The file must contain only data and no graphics, bitmaps or similar special structures which have nothing to do with the data
- All existing formulas must be replaced by numeric values
- The files must be "freshly" written, as multiple deletions and inserts into excel spreadsheets makes them unstable. The best method for this is: copy the data fields to the clipboard and "paste special" them as "values" in an entirely new excel spreadsheet without further content, and save this new excel spreadsheet with a new file name
- Additional information about each table may be entered in the cell A1 using the "comment" feature
- Additional information about the columns may be entered in the column headers using the "comment" feature
- The use of the "comment" function at other points is not allowed
- It is also inadmissible to define important specific features and abnormalities by cell formatting (alone). For example, to point out a period with erroneous data, delete these data instead of highlighting them by red coloring or the like. The cell formatting is lost in the processing of the data and must therefore not contain any significant information

ASCII-format

- Required column delimiter: semicolon. All lines must always have the same number of semicolons. Dots or commas as separators should never be used, since this may cause confusion with the thousands separator or decimal point.
- Alternatively, SPACE or TABULATOR can be used as column separators. The use of both separators together or alternately is not permitted and leads to uselessness of the files or extra costs. When using TABs, all lines must basically always contain the same number of TABs; there must only be one TAB between two columns. When using SPACES, several spaces can be used to separate columns, but only if no other separator than SPACE is used and no blank cells occur.
- Each row represents one record and all rows have the same number of columns (numbers)
- Empty cells (also missing values) are omitted when writing, so that in compliance with the specification (";" as the delimiter) a ";;" indicates the empty cell. Any numerical values which are supposed to indicate the blank cells (e.g. -99 999) are not permitted for semicolon separated files
- The comma serves as decimal point by default.
- Thousand separators may not be used.
- Any departure from the use of the comma as the decimal separator or the semicolon as a field separator has to be clearly documented.

Other formats

All other formats e.g. FoxPro cannot always be imported, so that considerable additional expenses may occur. Databases, provided that they are importable in principle, are subject to the same minimum rules as access databases.

If the data sets provided do not fulfil the demands described above, considerable additional expenses are to be expected.