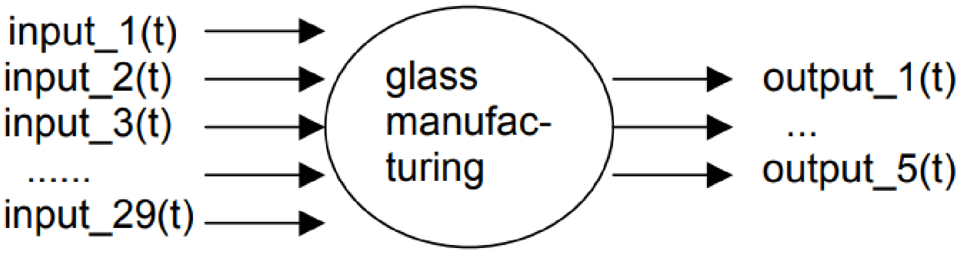
Wprowadzenie do Analizy Danych

Kierunek Systemy Teleinformatyczne, WE, sem. I

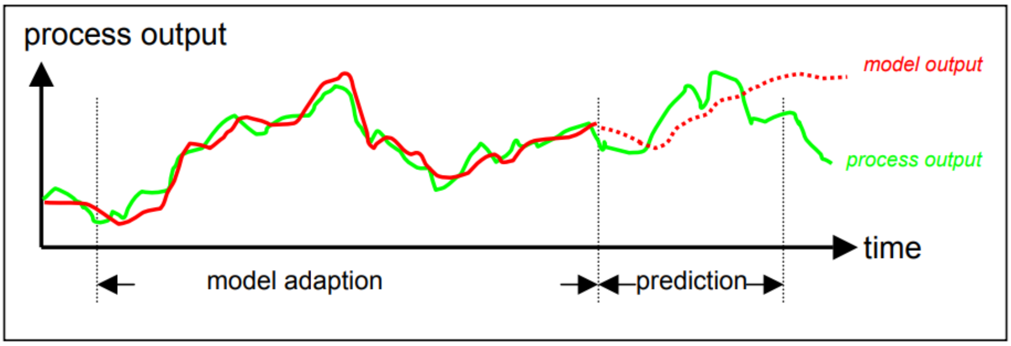
Laboratorium nr 11-12

Problem:

The quality of continuously manufactured glass products depends in a complicated manner on many input variables: there are influences which can be changed by the operator or by automatic controllers, and there are disturbances which can not be influenced. All measurable influences we consider as input variables for a mathematical multi-input-multi-output model describing relevant process variables (model outputs) representative for glass quality:



All inputs and outputs vary dynamically, and there might occur large time-delays: Changing an input variable may result in an output change starting only a couple of hours later and going on for up to several days.



Aim:

The aim is to find a mathematical model describing the relationship between 29 input variables and 5 output variables, all of them varying with time. Such a model probably has to be adapted to process state changes resulting from non-measurable influences. After adaptation to the current working point of the process the model should be able to predict the output variables over a certain time horizon, supposed the future behaviour of the input variables is known:

(The global trend of the outputs is much more important than the modelling of high-frequency variations.)

The criterion for model quality will be the closeness of model outputs to process outputs in a forecasting period of 1 weeks, where training data cover the preceding 11 weeks.

Description of the data file:

For model identification a data file (excel - **input_data_manufacturing_process.xls**; or ASCII - **input_data_manufacturing_process.txt**) containing (transformed) process data will be provided. No filtering, smoothing or feature selection of data has been applied, since data preprocessing is regarded as part of the problem. In the data file the participant will find 35 columns, where the content of the columns is as follows:

- column1: number of time step n
- column2-30: input data at time n
- column31-35: output data at time n.

The real meanings of input and output variables are strictly confidential and will not be communicated.

The data file consists of 8064 rows (time steps) - one data set every 15 minutes. This corresponds to process data over a period of 12 weeks. For the first 11 weeks (7392 rows) input and output data are given (training data set). For the last 1 weeks (672 rows) only input data are provided. These two weeks represent the forecasting horizon where the model quality will be evaluated (comparison with output data in this period which are not given to the competitor). Solutions will be ranked by accuracy of approximation of the output data during the forecasting horizon.

Specification of accepted solutions:

Solutions are expected in the following form:

1. Data file (excel) with 672 rows and 6 columns, where the rows correspond to time steps 7393 - 8064. The first column has to contain the number of time step, columns 2-6 must contain corresponding output data for outputs 1-5 predicted by the model using input data of the provided data file.
2. Report (pdf) containing:
    a. name of student(s)
    b. problem definition
    c. review of the approaches dedicated to the comparable problem (including review of the literature)
    d. description of the proposed approach chosen for modelling (data preprocessing, model structure, training algorithm)
    e. description of method of adaptation to current working point, if applied (splitting of training data in adaptation periods, adaptation algorithm, model forecasting quality on parts of training data)
    f. diagram(s) illustrating training results (comparison of model outputs with process outputs in training period or several adaptation periods) .