International Journal of Computers, Communications & Control
Vol. II (2007), No. 2, pp. 111-120

# Variable Selection and Grouping in a Paper Machine Application

Timo Ahola, Esko Juuso, Kauko Leiviskä

**Abstract:** This paper describes the possibilities of variable selection in large-scale industrial systems. It introduces knowledge-based, data-based and model-based methods for this purpose. As an example, Case-Based Reasoning application for the evaluation of the web break sensitivity in a paper machine is introduced. The application uses Linguistic Equations approach and basic Fuzzy Logic. The indicator combines the information of on-line measurements with expert knowledge and provides a continuous indication of the break sensitivity. The web break sensitivity defines the current operating situation at the paper mill and gives new information to the operators. Together with information of the most important variables this prediction gives operators enough time to react to the changing operating situation.

**Keywords:** variable selection, grouping, paper machine, web breaks

## 1 Introduction

Data-driven modelling requires always variable selection or grouping. In small systems, expert knowledge gives a clear basis for the variable selection since possible interactions and causal effects are known fairly well. In these cases, few modelling alternatives can be compared interactively. Variable selection becomes important when the number of variables increases, especially when normal process data is used. As a model should include a reasonable number of variables, a modular approach based on variable grouping provides a better process insight, which makes the model assessment easier.

In practical cases, variable selection is necessary either because it is computationally infeasible to use all available variables, or because of estimation problems when limited data samples with a large number of variables are present.

Variable grouping means finding feasible groups and combinations of variables for modelling. It is closely connected to data clustering since the interactions can depend on the operating area. In large-scale systems, the number of possible variable combinations becomes easily very large, e.g. the case models of the web break indicator included originally 24 variables, which mean 2024 alternative three variable combinations. The newest version has 72 variables leading to 59,640 three variable groups, 1,028,790 four variable groups and 13,991,544 five variable groups. Most of these alternatives are useless, and therefore, methods for selecting reasonable variables for modelling are crucial.

There is a lot of recent literature on variable selection and both model and data-based techniques are in use. Spectroscopic data, multi-sensor systems, multivariate analysis and modelling of large-scale systems seem to require efficient methods for variable selection. Four different methods for variable selection Ű genetic algorithm, iterative PLS, uninformative variable elimination by PLS and interactive variable selection for PLS - in partial least square (PLS) regression are studied and compared to a calibration made with manually selected wavelengths in [1]. The application is NIR analysis of pharmaceutical tablets. It has been found that multiresolution analysis (Haar wavelets) pre-processing before variable selection leads to simpler models with lower errors than single-wavelength selection in NIR data [2].

Wavelength selection for process monitoring has also been done using genetic algorithms (GA) coupled with a curve resolution method (OPA) [3]. Variable selection is also an important topic in using multiway methods in modelling NIR spectra from a pharmaceutical batch process [4]. NIR analysis of sugar cane juice has utilized partial least squares (PLS) pruning for variable selection [5]. UV-VIS and NIR spectrometry of oils takes advantage of the successive projections algorithm (SPA) in large-scale variable selection [6].

Quantitative structureŰactivity relationship (QSAR) studies require also sophisticated methods for variable selection. There is a report on applying multi-objective genetic programming (GP) to the HEPT data and constructing the nonlinear QSAR model using counter-propagation (CP) neural network with the selected variables [7]. Particle swarm optimizer (PSO) applies for the same purpose and the comparison with GP is in [8]. Norinder [9] reports also on the use of support vector machine (SVM) in QSAR. Statistical parametric mapping (SPM), relying on the general linear model and classical hypothesis testing, is a benchmark tool for assessing human brain activity using data from fMRI experiments [10]. Prediction-based variable selection has reported to give 82 % success rate in Quantitative StructureŰProperty Relationship models (QSPR) based on in vivo bloodŰbrain permeation data [11].

In multi-sensor systems, variable selection problem originates from two reasons: the high dimensionality in the data used is due to a high number of sensors or many features extracted, or both. Fuzzy ARTMAP classifier analyses the results from a 12-element gas sensor array [12, 13]. Fast wavelet transform is useful in feature selection before calibration in stripping voltammetry [14].

Principal component analysis (PCA) is a well-known method for variable selection. Testing of loadings and their estimated standard uncertainties are used to calculate significance on each variable for each component [15]. Variable selection can also mean identifying a k-subset of a set of original variables that is optimal for a given criterion that adequately approximates the whole data set [16]. The application of Principal Component Regression to the trajectories of the process variables (block-wise PCR) has given straightforward results without requiring a deep knowledge of the process [17]. In this case, variable selection methods and technical information of the process has allowed the process variables most correlated with the final quality be revealed.

Genetic algorithms (GAs) have been proposed recently for many applications including variable selection for multivariate calibration, molecular modelling, regression analysis, model identification, curve fitting, and classification. GAŠs are also incorporated with Fisher discriminant analysis (FDA) for key variable identification for trouble-shooting problems of the Tennessee Eastman process [18]. GA and simulated annealing have also been combined for reduction in the number of variables in neural network models [19]. Two other approaches for the selection of variables in neural networks are in [20] and [21].

This paper is organised as follows: Section 2 concerns with knowledge-based variable selection and grouping, Section 3 with variable grouping with data analysis and Section 4 with model-based variable selection. The case-based reasoning system for evaluating paper machine web breaks is shortly revisited in Section 5.

## 2 Knowledge-based variable selection

Knowledge can be used in decreasing the number of variables. For example, if we have a case with 10 process variables and group them in all possible groups with three, four and five variables, we end up to 582 groups. If we can, based on the process knowledge, include variable 1 in all groups with three variables, variable 10 in all groups with four variables, and variables 5 and 6 in all groups with five variables, we have 176 groups to analyse. This means that using process knowledge has decreased the number of alternatives by 70 percent.

Some variable combinations should be avoided, e.g. calculated variables should not be used together with the variables used in calculating them. Also a group containing a controlled variable and its setpoint is not usually appropriate. These problems are avoided by defining the inappropriate groups as non-groups, i.e. as variables groups, which should not be a part of any acceptable variable group.

## 3 Variable grouping with data analysis

Correlation is a statistical technique which can show whether and how strongly pairs of variables are related. Binary correlations and their combinations are used in pruning the set of acceptable groups defined by the domain expertise. For forecasting models, input variables should have a high correlation with the output variables, but a low one with each other. For case detection, causality is not always as clear: there is nor necessarily any definite output variable i.e. also groups where several variables have a high correlation between each other are acceptable. This sets new requirements for the model assessment.

In practical cases, the results from correlation analysis are improved with appropriate filtering and using correct time delays between the variables. Calculation of moving averages, medians and value ranges includes already a time delay, which depends on the calculation window and the applied methodology.

Nonlinear scaling is the essential feature in using Linguistic Equations method [22]. It improves the correlation analysis of curvilinear relationships, since the correlation analysis is a linear method. Finding patterns in data with high dimension is difficult. However, in data sets with many variables, groups of variables often move together as they are measuring the same phenomena. A host of clustering approaches helps in digging out these interactions.

As shown in Introduction, Principal Component Analysis (PCA) is a conventional method to decrease the dimensionality in data without losing the information stored in the correlated variables. It searches for new fewer linear combinations of the original variables that explain the most of the variance of the original data. These linear combinations can be viewed as a linear transformation to the hyperplane defined by the principal components or a rotation and a stretch that transform original data to a new bias.

Principal components are calculated by defining the eigenvectors of the covariance matrix or utilizing the singular value decomposition. Usually, only few first principal components (2 or 3) are used Ű they are enough to explain most of the variance in the data set. There are also extensions in the basic methods that apply for analyzing time trajectories.

## 4 Model-based variable selection

Isokangas and Ruusunen [23] describe the automated procedure for finding interactions between variables from large datasets. This occurs systematically by constructing simple dynamic model candidates with complete input combinations for data segments of the varying and sliding window size. The final analysis goes on according to the structure properties of the best candidate models.

Model candidate construction, validation and testing proceed in the following way: the half of all available data is used in training and validation so that model candidates are constructed systematically from the beginning of data with selected data window size. After a data window has been used for training, the window of the same size is taken for validation. The procedure uses a partly overlapping data window. For example, if the data window is 400 minutes, first models are constructed using training data from 1 - 400 minutes and data from 401 - 800 for validation of a model candidate under evaluation. Next, all model candidates are constructed using training data from 201 - 600 and validation data from range 601 - 1000 minutes. To define the right size of training data, different window sizes are systematically tested at this stage. Models are evaluated with the correlation coefficient and RMS-error measure using validation data. Best models are further tested with independent testing data, which is another half of available data.

# 5 Paper mill example

Paper web breaks commonly account for 2-7 percent of the total production loss, depending on the paper machine type and its operation. This could mean 1.5 million euros lost annually at a single paper machine. According to statistics only 10-15 percent of web breaks have a distinct reason. The most of the indistinct breaks are due to dynamical changes in the chemical process conditions.

The main area of interest in the indicator development is the paper making process before the actual paper machine. This includes also the short circulation and the wet end of the paper machine. In this area, the paper making process is typically non-linear with many, long delays that change in time and with process conditions, there are process recycles at several levels, there are closed control loops, there exist factors that can not be measured and there are interactions between physical and chemical factors. Also several different paper grades are produced with different production conditions and operating parameters.

This Section shows how to combine on-line measurements and expert knowledge in paper machine modelling in developing the sensitivity indicator for paper web breaks [24]. The indicator would give the process operators a continuous indication of the web break sensitivity in an easily understandable way. Being able to indicate the break risk would give a possibility to react on changes of the break sensitivity in time and therefore avoid breaks.

## 5.1 Experimental data

The actual measurements from a paper machine were used. The main interest was in paper machine variables and the variables just before the paper machine. The final selection of variables used expert knowledge and altogether 73 variables (72 variables + information on the break occurrence) were studied. These variables were supposed to influence on paper web breaks.

The measurements were collected from the mill automation system during normal operation and no special test runs were made. The measurements were used as such to retain their information content, and, on the other hand, to keep the application as simple as possible. Only a simple filtering was added to the indicator software to make rapid changes slower and to cut the outliers from the data. The measurement data was divided into periods of 24 hours. Further the data sets were classified into five categories, depending on how many breaks there were in one day: no breaks (0), a few breaks $(1-2)$, normal $(3-4)$, many breaks $(5-6)$ and a lot of breaks $(> 6)$.

## 5.2 Reasons for web breaks

Different statistical methods were used, but reliable correlations between single variables and web breaks did not exist. Therefore, the only way to proceed was the classification and modelling of break situations to find out differences between operating situations leading to breaks. Case-Based Reasoning was used for the identification of different operating situations instead of trying to predict a single break occurrence.

Identified operating situations contain information about how many breaks there will be in the near future and this information is given to the process operators as the web break sensitivity. The identification is performed using Linguistic Equation approach and Fuzzy Logic [24].

## 5.3 Correlation analysis

Before modelling, correlation analysis was used in order to find out binary interactions between different process variables. The basic tool used for these analyses was Microsoft Excel spreadsheet. The correlation exceeding 0.6 was considered worth mentioning. According to this analysis, correlation varies quite a lot with time. The variation in correlation rates is due to the usage of normal on-line measurements, which include the effects of different control operations.

The most important result of this analysis was that interactions vary in different operating situations, and the number of breaks also varies with time, and this was the basis for different case models. Due to different interactions, also different variables became important in different operating situations.

### 5.4 Model-based variable grouping

The web break sensitivity indicator was developed as a Case-Based Reasoning type application with Linguistic Equations approach and Fuzzy Logic [24]. The case base contains case models with different number of breaks. A new case is presented to the system as a collection of on-line measurements. The indicator compares the new case to the examples in the case base and uses the information of the best fitting case to calculate the predicted break sensitivity. As output the system gives numerical value for the predicted amount of breaks [24, 25]. Figure 1 shows the principal structure of the case base, and Figures 2 and 3 the different stages of Case-Based Reasoning.

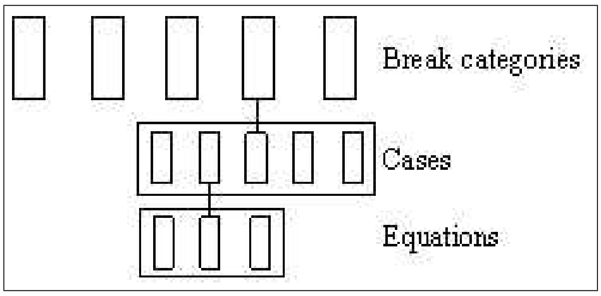

Figure 1: Structure of the Case Base [25].

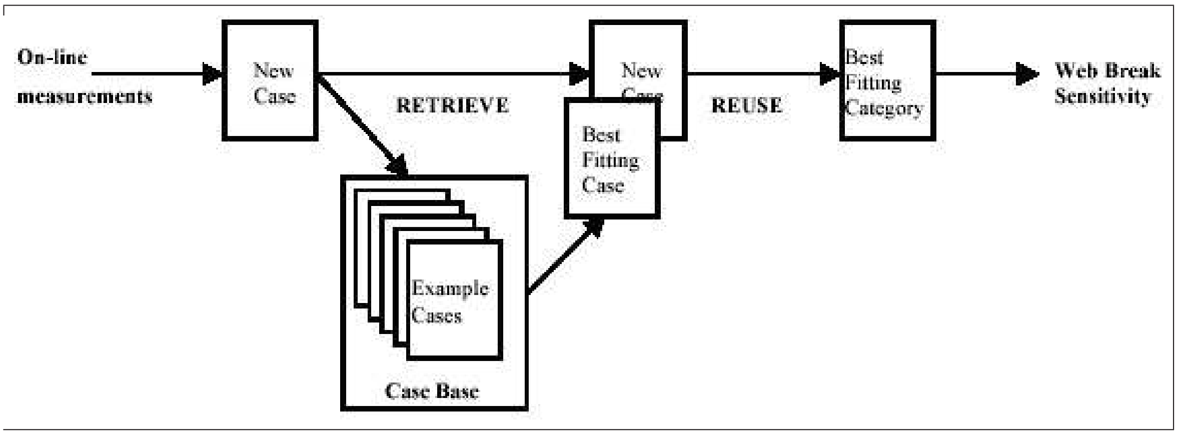

Figure 2: The structure of RETRIEVE and REUSE stages [25].

The Case Base of this application contains modelled example cases classified according to the number of related breaks. Models consist of equations that are stored as simple numerical matrices, which are indexed with break class information and number of examples in class. Equations itself describe the interactions between 3-5 variables. The variables in equations are found using a partly knowledge-based, partly model-based grouping technique.

For complex systems, a set of alternative variable groups are generated and models created with these groups. Process knowledge can be used in defining these groups. Another approach is to generate all possible groups containing three, four or five variables and modelling them. Groups can also contain different number of variables.

Correlation analysis has also use in grouping. It should be noted that for prediction the input variables should have a high correlation with the output variable, but a low one with each other. For state detection, causality is not clear, and the group where all variables correlate with each other are acceptable. Here, however, the limitations given in Section 2 must be taken into account.

Groups with three, four or five variables can be generated automatically with FuzzEqu Toolbox [22]. The generation of the alternatives is based on groups with three variables: all groups with four variables have one variable in common, and all groups with five variables have two variables in common. The subsets of the variables and the common variables in the groups with four or five variables can be based on process knowledge.

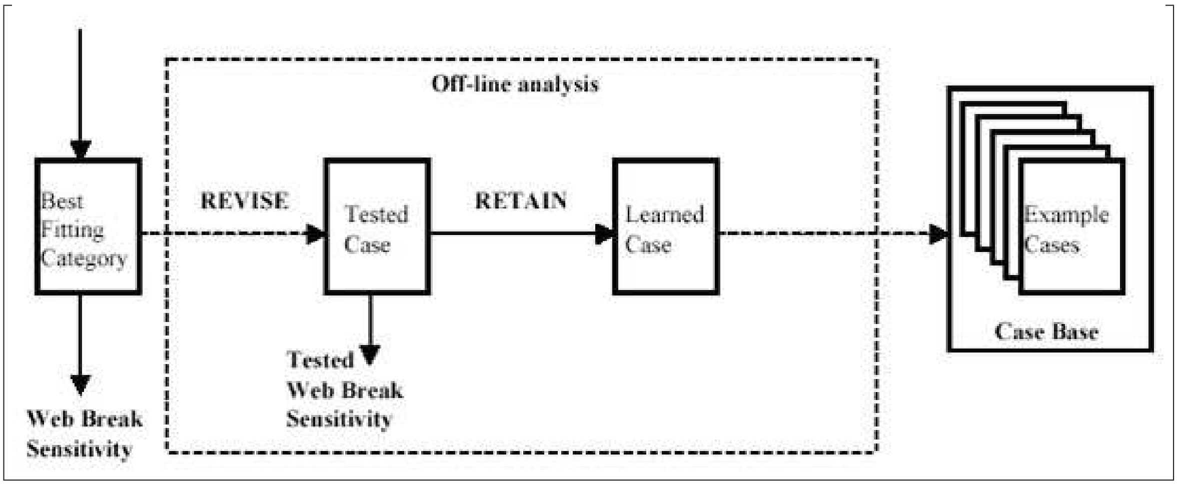

Figure 3: The structure of REVISE and RETAIN stages [25].

### 5.5 Some feedback Ű importance of parameters

After modelling, the importance of variables was analysed based on the occurrence of variables in case models. This was also considered as a useful tool to reveal less important variables not included in the models. For the operating personnel, the list of importance might give some new information about which variables are responsible for different operating problems. There is also some difference in the collection of important variables between cases with different number of breaks.

The user interface presents continuously the six variables that best describe the current operating situation. These are simply the variables of the two best fitting equations of the best fitting case. In addition to these, also the most important variable of the best fitting case model is presented with a trend value of 8 hours. The variables are presented with their current measurement values marked with colours as normal, low or high and very low or very high. This information gives the process operators useful information of the current process conditions together with the reason for the current break sensitivity level.

The follow-up of the variable importance can lead to the need to update the whole system. The system updating is a straightforward task, all though time consuming when the whole case base is changed. A single model with 73 variables takes only few minutes to build. The same time is required for validation and tuning and it makes altogether 15 to 20 minutes. The updating of the system with 40 cases could take the working hours of two days. However, the automation of these tasks makes this time shorter.

# 6 Summary and Conclusions

This paper has considered possibilities to variable selection in large-scale industrial systems. It introduced knowledge-based, data-based and model-based methods for this purpose. As an example, Case-Based Reasoning application for the evaluation of the web break sensitivity in a paper machine was introduced. The application was build with Linguistic Equations approach and basic Fuzzy Logic.

The Case Base of the system contains models of example cases with different number of breaks. A new case is presented to the system as a collection of on-line measurements. The indicator compares the new case to the examples in the case base and uses the information of the best fitting case to evaluate the break sensitivity. The latest version of the indicator operates with a case base of 40 example models. Although the size of this case base is rather small, the results have been considerably good compared to the real break sensitivity.

The indicator combines the information of on-line measurements with expert knowledge and provides a continuous indication of the break sensitivity. The web break sensitivity defines the current operating situation at the paper mill and gives new information to the operators. The web break sensitivity is presented as a continuous signal with information of the actual web breaks as a trend of 8 hours. The trend shows how the situation has developed and the current value gives the prediction for next 24 hours if the situation stays as it is now. Together with information of the most important variables this prediction gives operators enough time to react to the changing operating situation.

The variable selection and grouping utilize knowledge-based and model-based approaches. Automatic group and model generation makes also the interactive variable selection possible.

# References

[1] C. Abrahamsson, J. Johansson, A. Sparén, and F. Lindgren, Comparison of different variable selection methods conducted on nir transmission measurements on intact tablets, *Chemometrics and Intelligent Laboratory Systems*, Vol. 69, pp. 3–12, 2003.

[2] C.E.W. Gributs and D.H. Burns. Parsimonious calibration models for near-infrared spectroscopy using wavelets and scaling functions, *Chemometrics and Intelligent Laboratory Systems*, Vol. 83, pp. 44–53, 2006.

[3] S. Gourvénec, X. Capron, and D. L. Massart, Genetic algorithms (GA) applied to the orthogonal projection approach (OPA) for variable selection, *Analytica Chimica Acta*, Vol. 519, pp. 11–21, 2004.

[4] L. Stordrange, T. Rajalahti, and F.O. Libnau, Multiway methods to explore and model NIR data from a batch process, *Chemometrics and Intelligent Laboratory Systems*, Vol. 70, pp. 137–145, 2004.

[5] S.L.T. Lima, C. Mello, and R. J. Poppi, PLS pruning: a new approach to variable selection for multivariate calibration based on hessian matrix of errors, *Chemometrics and Intelligent Laboratory Systems*, Vol. 76, pp. 73–78, 2005.

[6] M.J.C. Pontes, R. Kawakami, H. Galvão, M.C. Ugulino Araújo, P.N. Teles Moreira, O.D. Pessoa Neto, G.E. José, and T.C. Bezerra Saldanha, The successive projections algorithm for spectral variable selection in classification problems, *Chemometrics and Intelligent Laboratory Systems*, Vol. 78, pp. 11–18, 2005.

[7] M. Arakawa, K. Hasegawa, and K. Funatsu, QSAR study of anti-HIV HEPT analogues based on multi-objective genetic programming and counter-propagation neural network, *Chemometrics and Intelligent Laboratory Systems*, Vol. 83, pp. 91-98, 2006.

[8] Q. Shen, J.-H. Jiang, C.-X. Jiao, G. Shen, and R.-Q. Yu, Modified particle swarm optimization algorithm for variable selection in MLR and PLS modeling: QSAR studies of antagonism of angiotensin II antagonists, *European Journal of Pharmaceutical Sciences*, Vol. 22, pp. 145–152, 2004.

[9] U. Norinder, Support vector machine models in drug design: applications to drug transport processes and qsar using simplex optimisations and variable selection, *Neurocomputing*, Vol. 55, pp. 337–346, 2003.

[10] M. Smith, B. Pütz, D. Auer, and L. Fahrmeir, Assessing brain activity through spatial bayesian variable selection, *NeuroImage*, Vol. 20, pp. 802–815, 2003.

[11] R. Narayanan and S.B. Gunturi, In silico ADME modelling: prediction models for bloodŰbrain barrier permeation using a systematic variable selection method, *Bioorganic & Medicinal Chemistry*, Vol. 13, pp. 3017–3028, 2005.

[12] E. Llobet, J. Brezmes, O. Gualdrón, X. Vilanova, and X. Correig, Building parsimonious fuzzy ARTMAP models by variable selection with a cascaded genetic algorithm: application to multisensor systems for gas analysis, *Sensors and Actuators B: Chemical*, Vol. 99, pp. 267–272, 2004.

[13] O. Gualdrón, E. Llobet, J. Brezmes, X. Vilanova, and X. Correig, Coupling fast variable selection methods to neural network-based classifiers: Application to multisensor systems, *Sensors and Actuators B: Chemical*, Vol. 114, pp. 522–529, 2006.

[14] M. Cocchi, J.L. Hidalgo-Hidalgo de Cisneros, I. Naranjo-Rodríguez, J.M. Palacios-Santander, R. Seeber, and A. Ulrici, Multicomponent analysis of electrochemical signals in the wavelet domain, *Talanta*, Vol. 59, pp. 735–749, 2003.

[15] F. Westad, M. Hersleth, P. Lea, and H. Martens, Variable selection in pca in sensory descriptive and consumer data, *Food Quality and Preference*, Vol. 14, pp. 463–472, 2003.

[16] J. Cadima, J. Orestes Cerdeira, and M. Minhoto, Computational aspects of algorithms for variable selection in the context of principal components. *Computational Statistics & Data Analysis*, 47(2):225–236, 2004.

[17] M. Zarzo and A Ferrer, Batch process diagnosis: PLS with variable selection versus block-wise PCR, *Chemometrics and Intelligent Laboratory Systems*, Vol. 73, pp. 15–27, 2004.

[18] L. H. Chiang and R.J. Pell, Genetic algorithms combined with discriminant analysis for key variable identification, *Journal of Process Control*, Vol. 14, pp. 143–155, 2004.

[19] A. Alexandridis, P. Patrinos, H. Sarimveis, and G. Tsekouras, A two-stage evolutionary algorithm for variable selection in the development of RBF neural network models, *Chemometrics and Intelligent Laboratory Systems*, Vol. 75, pp. 149–162, 2005.

[20] F. Dieterle, S. Busche, and G. Gauglitz, Growing neural networks for a multivariate calibration and variable selection of time-resolved measurements, *Analytica Chimica Acta*, Vol. 490, pp. 71–83, 2003.

[21] I. Drezga and S. Rahman, Input variable selection for ann-based short-term load forecasting, *Power Systems, IEEE Transactions on*, Vol. 13, pp. 1238–1244, 1998.

[22] E. K. Juuso, Integration of intelligent systems in development of smart adaptive systems, *International Journal of Approximate Reasoning*, Vol. 35, pp. 307–337, 2004.

[23] A. Isokangas and M. Ruusunen, Systematic approach for data survey, in *Proceedings of the International Conference on Informatics in Control, Automation and Robotics. September 14 - 17, 2005, Barcelona, Spain*, pp. 60–65, 2005.

[24] T. Ahola, Intelligent estimation of web break sensitivity in paper machines. Doctoral dissertation. University of Oulu, Department of Process and Environmental Engineering. Acta Universitatis Ouluensis, Technica C 232, 92 p., Oulu, 2005.

[25] T. Ahola and K. Leiviskä, Case-based reasoning in web break sensitivity evaluation in a paper machine, *Journal of Advanced Computational Intelligence and Intelligence Informatics*, Vol. 9, pp. 555–561, 2005.

Timo Ahola, Esko Juuso, Kauko Leiviskä
University of Oulu, Control Engineering Laboratory
P.O. Box 4300, FI-90014 University of Oulu, Finland
E-mail: esko.juuso@oulu.fi
Received: March 21, 2007

Timo Ahola, born in Hämeenkyrö, Finland on December 18, 1965, recieved his M.Sc. (Eng.) in Process Engineering in 1992 and Lic. Tech. in Control Engineering in 2001 from the University of Oulu. He recieved D. Tech. in Control Engineering from the University of Oulu in 2006 with the thesis on intelligent estimation of web break sensitivity in paper machines. He worked as a research scientist in the Control Engineering laboratory at the University of Oulu 1993-2006. Since 2007 he has been belonging to the Outokumpu Stainless Oy, Tornio Research Centre, Finland. Presently he is a research engineer in process development and he is working with predictive maintenance issues.

**Esko Juuso**, born in Ylitornio, Finland on December 12, 1951, received M.Sc. (Eng.) in Technical Physics in 1979 from University of Oulu. He has worked as research engineer in Outokumpu Metallurgical Research Centre and computer analyst in Outokumpu Electronics. Since 1986, he has been belonging to University of Oulu, Oulu, Finland. Presently he is Senior Assistant in Control Engineering. He has been active in Finnish Simulation Forum (FinSim), Scandinavian Simulation Society (SIMS) and EUROSIM, currently he is chairman of FinSim. He has been member of Steering Committee and co-chairman of Technical Committee on Production Industry in EUNITE Network of Excellence, 2000-2004. His main research fields are intelligent systems and simulation in industrial applications, including control and fault diagnosis. In 1991 he introduced the linguistic equation (LE) methodology, and he has authored more than 200 publications on his research field.

**Kauko Leiviskä**, born in Pyhäntä, Finland, 1950, received M.Sc.(Eng.) in Process Engineering in 1975 and Lic. Tech. in Control Engineering in 1976 from the University of Oulu. He received the D. Tech. in Control Engineering from the University of Oulu in 1982 with the thesis on short term production scheduling of the pulp mill. Since 1975, he has been belonging to University of Oulu, Oulu, Finland. He has been Professor of Control Engineering and Head of Control Engineering Laboratory in the same University since 1988. He has been active in IFAC since 1988, currently he is member of IFAC TC on Large Scale Complex Systems, TC on Cognition and Control and TC on Mining, Mineral and Metal Processing. He has been member of Steering Committee and chairman of Technical Committee on Primary and Process Industries in ERUDIT Network of Excellence, 1997-2000, and the scientific director of EUNITE, the European Network of Excellence, 2000-2004. He is participating in EU/CA project NISIS (Nature-Inspired Smart Information Systems). A list of more than 200 publications of which he is (co)author is available. Recently his work concentrates on modelling and control of industrial processes, intelligent control methods, production scheduling and millwide control. He has also been consulting industry on control engineering and millwide control applications.