ADVANCED TECHNOLOGIES FOR NEW PARAMETERS AND MEASUREMENT CONCEPTS

M. JAUZEIN, Director of the GIP Stelor and Research Director of IRH Environnement, 11 bis, rue Gabriel Péri, 54 500 Vandoeuvre, France

1. Introduction

The development of advanced technologies for data treatment in the field of computer science and numerical tools gives a sound basis for information structuring, analysing and drawing and for the understanding, modelling and validation of parameters relationships.

The monitoring of water quality can be performed through a large set of classical parameters, innovative highly specific measurements, and sensitive sensor signals. But then the number of individual data and the interpretation of less specific informations become new limiting factors. To overcome these limiting factors, advanced technologies for data treatment can be optimised for the derivation of virtual or integrated parameters which are the keys of observed time or spatial behaviours.

The phenomenological approach is generally used for identifying relationships between variables on the basis of deterministic models describing elementary phenomena steering studied behaviours (thermodynamical laws, mass balance constraints, kinetic laws, ...). In the case of physico-chemical parameters describing the quality of water, these relationships are generally complexes and often non linear including additional parameters that are not known only on a statistical basis or with poor confidence. This general approach gives many rules for understanding multivariate behaviour of real systems in surface water and groundwater monitoring, pollutant transport in natural waters, drinking water or sewage collection networks, and water treatment processes. Nevertheless, the deterministic approach is limited by the lack of confidence for some additional parameters due to insufficient data basis or uncontrolled parameter fitting.

The " black box " approach is an alternative to this phenomenological approach. Without a clear knowledge of the phenomena steering the multivariate behaviour of real systems, it consists in deriving key parameters from a large set of data characterising a studied object. The proposed data treatment strategy has the aim of reducing the set of data to an optimised set of integrated or virtual parameters describing the main shapes of the multivariate systems. Multivariate statistical analysis is one of the more classic approach that has been applied in this context. This first pathway for solving the problem is limited by the fact that only linear correlations are analysed. The non linear relationships need to state preliminary data transformations using empirical or phenomenological non linear laws. Other pathways are now open using, for example, virtual connections through neural networks or fuzzy logic rules.

This paper has not the aim of giving a theoretical exhaustive view of those advanced numerical techniques but to illustrate their potential use in the field of water quality monitoring.

2. State of the art

The multivariate linear statistic analysis is one of the most classic way to reduce information packages to principal factors or components. Thus, it is not necessary to remind those techniques, but only to outline the limits of this approach for taking into account the existence of non linear relationships between variables.

2. 1. Neural networks tools [1 to 11 and 24]

The human brain is able to memorise informations, to classify them, to relate and identify them. Through the observation of the biological mechanisms governing brain functions, several fundamental principles has been identified which allow to reinforce existing techniques and develop innovative data treatment methods. These tools are based on a neuromimetic or connectionist approach. As a mater of fact, neural networks try to mimic neural structures through connection networking. The high number of connections, the non linearity, the relationships between input and output parameters are the main characteristics of those numerical tools. The connectionist term is more appropriate for outlining the distance between present neural networks and the biological model.

In the present state of the art, connectionist methods are not universal solutions, even they can be applied successfully in many practical applications. In the future, the combination of neural networks with expert systems, classical algorithms and complementary tools will generate the more interesting data treatment technologies.

The basic element of a neural network mimic the cell of a brain system. The characteristic functions of these cells are the reception of signals from adjacent cells, the integration of these signals, the production, transport and delivery of a new signal to other cells. Without giving more details on the biological model, it is important to remind that the cell integrates continuously the input signals through synapses connections. To mimic these basic characteristics, the main idea is to use basic elements connected in networks through Artificial Neural Networks : a graph where nodes are elementary cells and arrow are connections. Cells can be entirely networked or structured in several layers.

The layered neural networks can be compared to statistical multivariate analysis. Specific neural network structures are exactly performing Principal Component Analysis or Discriminent Factorial Analysis. The main advantage of neural networking is to allow the integration of non linearity in the optimisation process through the activation function transforming the input function in a different output function. If preliminary non linear transformations can be applied before statistical multivariate analysis the problem of selecting them is unsolved. For neural networks, the selection of simple non linear activation functions makes the design of the data treatment easier.

But, the classic multivariate analysis gives more interesting results in terms of confidence analysis and correlation indices compared to present neural network tools. Another limiting factor is the selection of the optimal structure for the neural network. A strict optimisation strategy is generally too time consuming and a empirical approach is applied.

These ANN tools are directly applied to data reduction problems. In the case of time dependent series, it is possible to use as input data the multivariate signals for different past time data series to estimate an output data set at the present state. Similarly for space dependent data, there is potential for developing interpolation strategies based on the use of nearest multivariate data set to estimate an output data set on a specific location.

2. 2. Fuzzy logic tools [12 to 23 and 24]

The fuzzy logic approach, introduced by L.A. Zadeh in 1965, allows to treat informations easily with poorly defined concepts. This approach mimic human cognition processes. In addition, the concept of fuzzy object allows to define the link between this object and a well defined category of objects. For example, the limit between cold and hot temperatures refers to individual sensations which can differ between persons. The relationship between the words cold or hot and the personal feeling of each individual from a population includes a significant level of imprecision. This fuzzy limit between the category of situations cold and hot needs to be quantified. The fuzzy logic approach consist in establishing rules for linking the " hot " domain to numerical rules such as " temperature from 40 to 100°C ". In the case of misunderstood phenomena, this approach can be interesting compared to the deterministic one. The fuzzy definitions is a basic characteristic of human cognition. In many situations, fuzzy informations, fuzzy rules of interpretation and decision are used by human beings. Thus, the fuzzy logic tools allows to combine fuzzy data and precise numerical data in the process of data treatment and related decision making. Compared to conventional numerical limits, it also allows to take into account their imprecision through formalised precise rules. As a consequence the fuzzy logic concept is based on a precise numerical methodology to describe fuzzy systems.

A fuzzy set of objects is a class of object characterised by a continuous inclusion degree law ranged between 0 and 1. Compared to classic sets of objects where the inclusion law is binary (0 for excluded objects and 1 for included objects), this concept is more flexible. Fuzzy rules has been introduced by Mandani in 1974 for monitoring real systems. They are based on " if ... then ... " logic algorithms and are potential formalisms for developing expert systems on automated cognition processes. The main difficulty of this approach is to define the number of necessary rules which can be wording rules or functional rules (numerical law s).

The use of fuzzy logic for describing multivariate systems has appeared more recently It's a complementary tool for studying non linear or complex multivariate situations. These numerical tools are precision and complexity limited but their optimal design is very efficient for qualitative analysis of real systems. As neural network approach, it allows to derive output key variables from complex input data sets without a clear phenomenological description. Recently, a theoretical study of fuzzy logic systems has shown that they are universal

estimators or mono-variate functions and efficient estimators of multivariate functions. Some publications are also dedicated to optimised decomposition of linear system in fuzzy sub-systems. Finally, the design of non linear relationships is easier through fuzzy approaches. In addition, the fuzzy approach is available for interpolation procedures in the case of time dependent series and thus potentially for space dependent variables.

To perform the initial fuzzy analysis of a system, it is necessary to transform the variables in the form of fuzzy ones : this stage is called the fuzzyfication of the problem. Then the variables are treated with fuzzy logic rules and then obtained fuzzy results are transformed in the form of numerical data results : this last stage is called the defuzzyfication.

3. Monitoring of a water treatment plant with time dependent series

3. 1. The water treatment process

The water treatment process which has been studied is a classic multistage activated sludge process including oxygen injection systems for the activation of biological degradation phenomena [25]. It includes three identical bioreactors used for the continuous flow treatment of waste water fluxes. Each bioreactors can be steered for activating specific degradation processes depending on the quality of the waste water which is time dependent on an unpredictable way. Monitoring systems are installed for the following of pollution parameters and to control the treatment process.

3. 2. The data available

A specific set of parameters has been selected for the case study. It includes measurements of specific pollution parameters, biomass monitoring parameters, physical parameters (temperature and flow rates) and process control parameters (oxygen and biomass increase). Due to confidential reasons, the case study is presented with letters for identifying parameters (A to K for input parameters and L to N for output parameters).

The main set of variables are characterised by a cyclic evolution. The variables A to G show similar time behaviours but the sampling period differs for each parameter.

3. 3. Comparison of three advanced data treatment strategies for monitoring the water quality

Three methods have been used for time dependent series which has been preliminary normalised on the basis of the average, the maximum and minimum values :

- the auto-regressive approach using input data (AR) or a moving average approach including output data treatment (ARMA),
- the neural networks approach,
- the fuzzy logic approach.

The auto-regressive approaches are based on multi-linear regression tools applied to finite time series of input data (AR) or finite time series of input and output data (ARMA). The implementation of this approach needs to define a learning data base, a validation data base, a design of individual model structures, the fitting of parameters on the learning database for each model and the selection of the best model using a single criteria applied to the use of the model on the validation database. For this type of numerical tools, the ARMA approach is more efficient then the simple AR procedure for the same number of additional parameters.

The neural network approach has been applied through the test of three layers systems : one input layer, one intermediate layer and one output layer. The input set of data can include input and output past time series. The selection of input data sets has been performed on the basis of previous results obtained with the ARMA approach. The main results obtained is that, for the same number of parameters the results obtained with neural networks are better than using classic ARMA methods. This seems to be related to the easiest description of non-linear behaviours with neural networks than with multi-linear regression analysis.

The test of fuzzy logic approaches has been performed on only one input data and one output data. The main results obtained are a good estimation though learning runs, a high number of rules (about two thousand) which needs a prioritisation strategy, and a failure in the validation process.

This comparison of three type of advanced techniques for time series data treatment on a real system has given interesting preliminary results. Using present and past input data and sometimes past output data, these methods can be used for the prediction of key parameters for the monitoring and control of waste water treatment processes. The preliminary results obtained show that :

the inclusion of past output data jointly with present and past input data gives better results for the same number of input parameters.
the Auto-Regressive Moving Average and the Artificial Neural Network approaches are efficient compared to the fuzzy logic approach.
the Artificial Neural Network seems to be more adapted to this specific application on time data series.

A sequential use of ARMA and ANN is adapted to the multivariate analysis of time dependent behaviour. A more complete use of present input and past input and output data with this sequential approach gives a sound basis for future developments. Additionally, the development of moving learning procedures can be studied for on line correction-prediction strategies adapted to process phenomena evolution. Concerning fuzzy logic approaches, the fuzzyfication process has to be optimised to generate a limited number of key rules which will be easier to interpret and manage for deriving interesting data treatment alternatives.

3. 4. Test of a multi-model approach using fuzzy logic

A multi-model structure has been develop on the basis of a fuzzy logic approach to study the multivariate behaviour of the tested system. A specific formalism generalised from " Hammerstein " has been tested and a specific parameter fitting techniques has been developed. This formalism includes a finite number of rules based on non-linear static models coupled with a single linear discrete transfer function for the dynamic part of the relationship.

The combination of simple models which are efficient within a small range of process running conditions is designed through a specific aggregation rule able to give an algebraic relationship between input data to each output parameters on the whole range of process running conditions.

The fuzzification of each parameter is performed as a preliminary data treatment and consists in a partitioning of numerical data. The establishment of fuzzy rules is performed through a multi-model structuring of the relationship between input parameters and each output parameters. Then, the defuzzification is processed through classic numerical procedures giving a result for each elementary model rules. Then an aggregation law is used to derive the global estimation of each output parameters.

The developed identification procedure for parameters is based on conventional non-linear equation iterative solving (Gauss-Newton or Levenberg-Marquardt methods) applied to the formalism generalised from " Hammerstein ".

For the application of this approach to the presented case study, normalised data has been employed. Then some exotic values has been eliminated and a moving average filter has been applied to the original data set. The selected output data for illustrating the method is the UV absorption parameter.
(see figure 1 and figure 2)

The first step is performed through the optimisation of monovariate analysis between each input parameter and the studied output parameter. A prioritisation of obtained correlations is obtained and allows to perform a step by step identification of multivariate descriptions. In comparison with classic multivariate linear statistic analysis, this techniques allows to classify the interest of each variables as principal components and select the number of parameters on the basis of the increased correlation coefficient. From a correlation coefficient of about 0.88 obtained with a monovariate model, it increases up to 0.98 using all the available information. But three variables are sufficient to obtain a correlation coefficient close to 0.94 (see figures n° 1 and 2). The three variables multi-model has been analysed in terms of structure of the " Hammerstein " generalised model and sensitivity to each parameter.

The study of the dynamic part and its influence on the correlation coefficient indicates that the last point in the time series is optimal giving the maximal correlation and that the optimal order for this dynamic part of the model is 1. Then the number of rules has been studied giving 3 as an optimal number just through graphical determinations.

This learning procedure applied to one monitoring campaign has been validated on the basis of two other campaigns. For one of these campaigns, the results are promising and for the other, characterised by higher perturbations, the results are less convincing due to a transient drastic change in the quality of waste waters.

This study outlines the efficiency of the multi-model fuzzy logic approach for the estimation of multivariate and non linear behaviour of real systems. Some enhancements are needed for the optimisation of the multi-model structure. Lastly, one can note that the UV absorption parameter is the expression of the residual organic matter content in the effluent which can be related to the chemical oxygen demand (COD) and used for a feed-back control of the process.

Figure 1 : Simulation results for B (UV) obtained with 1 input parameter (G)

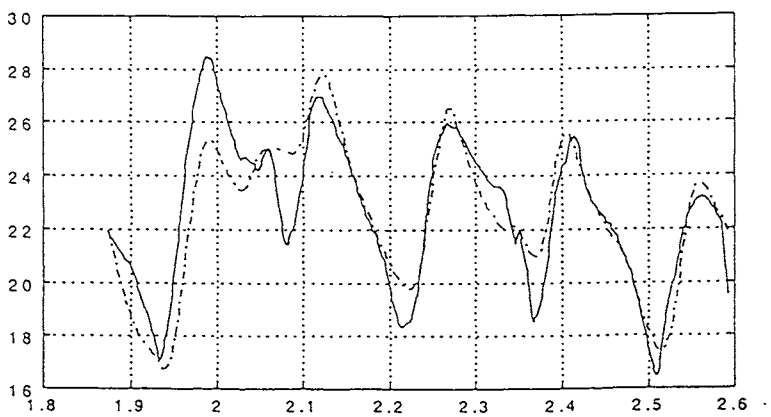
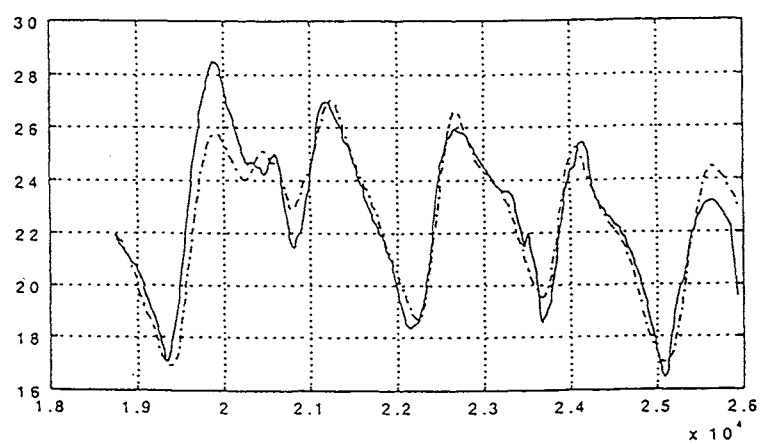


Figure 2 : Simulation results for B(UV) obtained with 3 input parameters (G, F and B)



4. Monitoring of groundwater quality with space dependent series

4. 1. The groundwater pollution source

Waste deposits are one of the pollution sources generating an acute or chronic degradation of water resource quality. For the assessment of contaminated sites, it is necessary to obtain maps giving a clear idea of the space distribution of contaminants. The selected case study is a waste deposits containing a mixture of municipal and industrial wastes [26]. A large range of organic solvents has been detected in the leachates of this site. These solvents are mainly mono-aromatic compounds like toluene (BTEX compounds) and halogenated aliphatic compounds (HOC compounds).

4. 2. The data available

In the case of the presence of volatile organic compounds (VOCs), it is possible to use global sensors delivering data on the presence and concentration of contaminants in the soil and subsoil gas phase. The main advantage of this sensors is the possibility to design field screening campaigns based on a large number of localised measurements obtained on the basis of a regular grid. The main limit of these sensors is the delivery of a global signal from each sensor, which is generally highly sensitive but only poorly selective. To overcome this problem, the use of several sensors at the same time can be proposed for delivering a multivariate signal which can exhibit a better selectivity. Another limiting factor for the use of this screening techniques is the indirect relationship with real groundwater contamination. As a matter of fact, even the partitioning of VOCs between the gas and the aqueous phase is linear, the relationship between the measured signal and the groundwater contamination is not necessarily linear. In addition, the presence of VOC mixtures with a large range of physico-chemical properties generates non-linear dependencies for global measurement systems.

To obtain significant maps of the contamination of groundwater by organic solvents, the following methodology has been applied : combined measurements with three VOCs global sensors on about 80 localised points, space dependent data treatment of results, selection of about 25 contrasted points, sampling and analysis of groundwater at this selected points, correlation analysis with previous results for estimating the relationship between VOCs mapping parameters (input data) and groundwater regulation parameters (output data).

The VOC sensors are a photo-ionisation detector, a semi-conductor sensor and a lipidic membrane piezoelectric sensor. The measurements have been performed using the installation of fixed perforated tubes down to a 3 metres depth. The gas phase is equilibrated with the subsoil gas phase during more than one day in the closed tubes. Then the gas phase of the tubes are characterised with the three sensors. Concerning aqueous phase sampling and analysis, the same tubes has been used for leachate sampling and the analysis of a set of priority pollutants have been performed through gas chromatography and specific detection of individual VOC.

4. 3. The methodology developed for the space dependent data treatment

Due to the fact that each detector gives a complementary response which is partially redundant with the others, a classical multivariate statistical analysis can give a better idea of the data redundancy determining principal components and their contribution to the global variance of the system. Consequently, this approach has been followed to identify new virtual parameters using a Principal Component Analysis (PCA) on the basis of normalised data sets. The result of this data treatment is the prioritisation of three virtual parameters with the help of variance explanation indices. Compared to initial parameters, these virtual parameters are characterised by their orthogonal behaviour (and gaussian behaviour if possible) for the studied data set.

Then a set of maps can be obtained for each virtual parameters to visualise the main characteristics of spatial heterogeneities (see figure 3). The proposed method is based on kriging after a variographic analysis of space dependent heterogeneities. The same type of model variograms (spherical) with different additional parameters (nugget effect, scale and maximal variance) has been derived from experimental data. Then these model variograms are used for interpolating data and delivering kriging maps of virtual parameters distribution.

4. 4. Correlation study with conventional groundwater analysis

The combination of field VOCs measurements (virtual parameters) and laboratory results on a single set of 25 localised points has been studied with a principal component analysis. Consequently, specific individual or groups of analytical variables can be correlated with the identified virtual parameters. Using the first three principal components and the three virtual parameters, it has been possible to determine optimal multi-linear models to transform virtual parameters (input data) into estimators of specific or groups of analytical variables (sum parameters). A parallel combination of the characteristic weight of the virtual parameters in the PCA allows to determine a confidence criteria for each multi-linear model. For example, a specific individual components highly correlated to a virtual parameter with a high weight can be estimated with a high confidence criteria (close to 1 as for benzene in figure 4). On the contrary, if the parameter to be estimated is only correlated with a virtual parameter characterised by a small weight, the confidence level will be too small (close to 0).

This specific study is an illustration of the combination of multivariate and space dependent statistical analysis in the field of water quality monitoring. The tools that has been tested are normally devoted to gaussian variables and linear relationships between variables. Thus, the development of alternative approaches based on neural networks or fuzzy logic for the studying of space dependent variables could be promoted. In addition, the proposed approach outlines the interest to develop multi-variate sensor systems based on spectral signals or multi-sensors arrays and the crucial need of a validation stage for deriving confidence criteria of produced maps.

Figure 3 : Maps obtained using the first principal component as virtual parameter

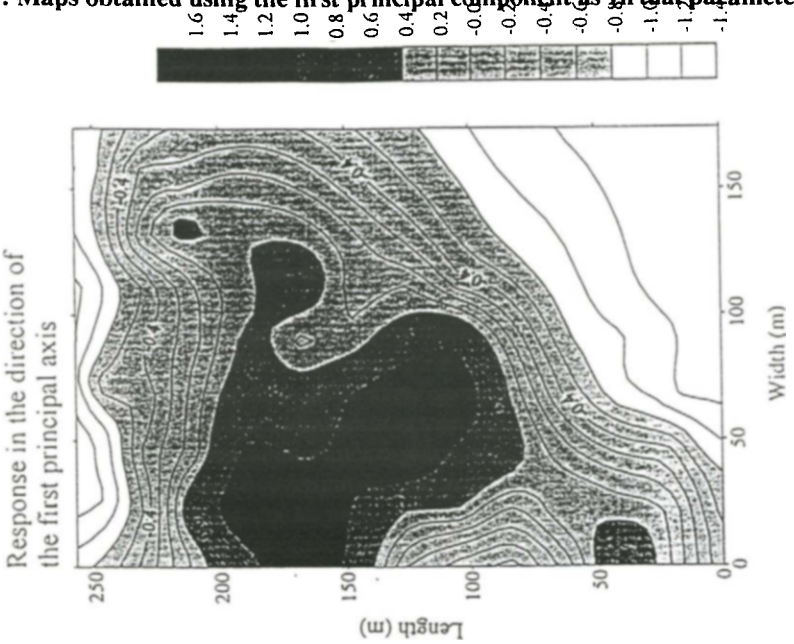
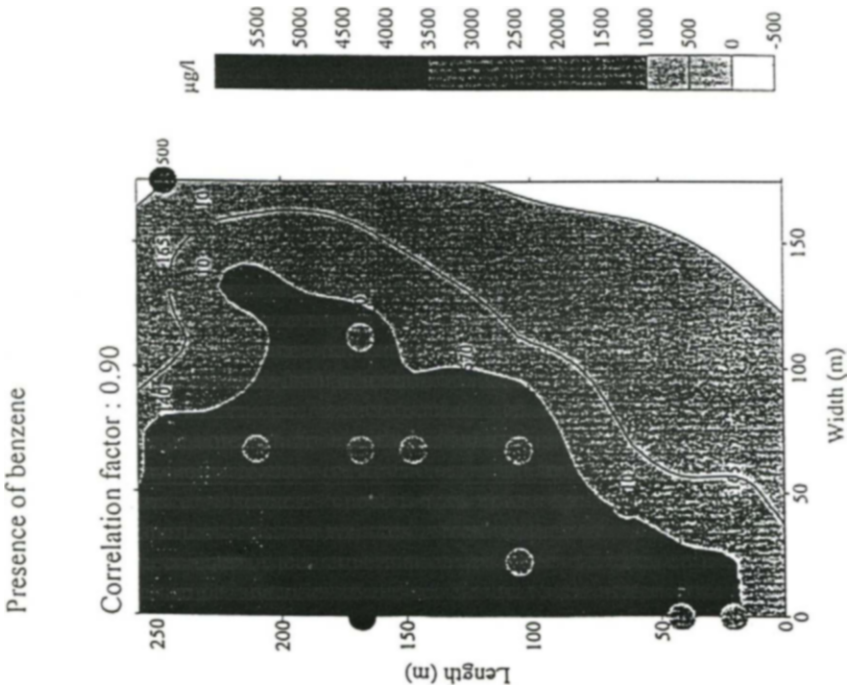


Figure 4 : Comparaison of a benzene concentration estimation and measured concentrations in leachates with a 0.90 confidence indice.



5. Conclusions And Perspectives

Three main types of chemometric multivariate techniques has been presented :

classic linear statistical analysis including multivariate statistics, multi-variate regression approaches for time dependent variables and geostatistical approaches for space dependent variables,
neural network approaches,
fuzzy logic approaches,

These advanced technologies for data treatment has been illustrated through case studies concerning a waste water treatment process and a contaminated groundwater plume. The main conclusion of this scientific and technical note is that the presented tools are now available for specific demonstration in the field of water quality monitoring either for time dependent or space dependent parameters. The range of application fields is large and the parallel development of sensor arrays of spectral signal sensors will increase the interest of these efficient numerical tools. To derive virtual or integrated parameters directly available for decision making from large and complex data sets, the optimal combination of phenomenological, statistical, neural and fuzzy approaches seems very attractive for future research and development in the field of measurement and testing or in the field of advanced technologies.

6. Acknowledgements

I acknowledge F.D. RAVOIRE, J.Y. CATROS, V.QUELIN and E. RANNOU from THOMSON-CSF, Services Industrie, J. RAGOT and C. LEVERINI from the CRAN-CNRS of Nancy (Centre de Recherche en Automatique de Nancy) and F. COLIN and G. GRAPIN from IRH Environnement in Nancy for their important contribution to this paper as partners in a joint research programme financed by the Ministry of the Environment (DGAD/SRAE/94059) on " black box " modelling approaches. I also acknowledge P. ROCKLIN from the NANCIE (International Centre for Water of Nancy) for its PhD contribution in a specific field application financed by the Ademe (National Agency for the Environment and Energy Management). Finally, I want to outline that a part of this work has been founded by the " Environment and Climate " european research programme (DGXII) in the field of " technologies for the environment protection " for the measurement and modelling of the VOCs mobility in soil and groundwater systems.

REFERENCES

1. E. Davalo and P. Naim, Editions Eyrolles, deuxième Edition, Des réseaux de neurones (1990).

2. C. Muller and M. Radoui, XXIIème Journées de Statistiques Tours, Réseaux neuromimétiques et analyses des données, Electricité de France (E.D.F Clamart) (1990).
3. R. Sobral, S. Canu, 71ème Congrès des AssociationsGénérales des hygiénistes et techniciens municipaux Annecy 15-19 avril 1991, textes de conférences pp 399-409, Application des réseaux de neurones artificiels à la prévision : la consommation d'eau (1991).
4. S. Thiria, LRI Université de Paris XI and Cedric CNAM, Réseaux de neurones : liens avec les techniques statistiques (1991-1992).
5. R. Tomassone, E. Lesquoy and C. Mullier, Inria Masson, page 24, La régression nouveaux regards sur une ancienne méthode statistique (1983).
6. F. Badran, S. Thiria, F. Fogelman Soulie, Kadratoff et Diday Editeurs, Comparaison analyse des données et réseaux multicouches : Introduction symbolique et numérique à partir de données (1991).
7. H. Bourlard, Y. Kamp, Manuscript M217 Philips Research Lab, Auto-association by multilayer perceptrons and singular value decomposition, (1987).
8. H. Bourlard, Y. Kamp, in IEEE 1st ICNN San Diego, Multilayer perceptrons and automatic speech recognition, (1987).
9. H. Bourlard, C.J Wellekens, Philips Research Laboratory Belgium, Links between markov models and multilayers perceptrons (1989).
10. G. Dreyfus, L. Personnaz, Journées Internationales des Sciences Informatiques Tunis, Intriduction au réseaux de neurones formels (1990).
11. P. Gallinari, F. Fogelman Soulie, S. Thiria, 2nd annual international conference on neural network classifiers San Diego, Multilayer perceptrons and data analysis "Neural Networks for computing" (1988).
12. M.B Beck, IIE Proceedings Vol. 133, Pt. D, n° 5, Identification, estimation and control of biological waste-water treatment processes" (1986).
13. J.C Bezdek, Vol 1 n° 1 pp. 1-6, "Editorial - fuzzy models - What are they and why ?", IEEE Transactions on Fuzzy systems (1993).
14. G. Block, Spécialité Automatique, Modélisation des procédés : des méthodes, un outil logiciel", Thèse - Doctorat de l'Université de Nancy I (1988).
15. D. Filev, Vol. 5 pp 281-290, Fuzzy modeling of complex systems, International Journal of Approximate reasoning (1991).
16. B. Kosko, IEEE Int. Conf. on Fuzzy Systems, San Diego pp. 1153-1162, Fuzzy systems are universal approximators.
17. L. Ljung, Prentice Hall, System identification. Theory for the user (1987).
18. E.H. Mandani, Proc. IEE, vol. 121 pp. 1584-1588, Application of fuzzy algoritms for control of simple dynamic plant (1974).
19. M. Sugeno, G.T Kang, Fuzzy sets and systems, 28 pp. 15-33, Structure identification of a fuzzy model" (1988).
20. T. Tagaki, M. Sugeno, IEEE Transactions on systems, Mann and Cybernetics, 15, pp.116-132, Fuzzy identification of systems and its application to modelling and control (1985).
21. E. Tan, G. Mourot, D. Maquin, J. Ragot, CNRS URA 821, Identification of fuzzy models, Centre de Recherche en Automatique de Nandy (1994).
22. L.X WANG, IEEE Int. Conf. on Fuzzy Systems, San Diego, CA pp.1163-1170, Fuzzy systems are universal approximators (1992).
23. L.A Zadeh, Information and Control, Vol. 8 pp 338-353, Fuzzy sets.

24. M. Jauzein, Gip Stelor RH 96-08 Ministère de l'Environnement, Rapport bibliographique (1996).
25. M. Jauzein, Gip Stelor Ministère de l'Environnement, Rapport final en cours d'édition (1997).
26. P. Rocklin, Thèse pour présentation de doctorat INPL, Conception de sondes de détection et dévleoppement de techniques de mesure in situ de la contamination de matrices solides par des composés organiques volatils (1996).