Porto PowerTech'2001,
IEEE Power Tech Conference ,Porto, Portugal, September 2001

Application of Data Mining Techniques to Identify Structural Congestion Problems under Uncertainty

E. F. Sánchez-Úbeda, J. Peco, P. Raymont, T. Gómez, *Member, IEEE*, and S. Bañales, A. L. Hernández

***Abstract*-- This paper proposes a novel methodology to identify congestion problems under both "traditional" and "new" uncertainties such as generation costs, location and size of new generators, retirement of old ones, generation patterns, etc. The methodology allows not only identifying the transmission paths and corridors which will have congestion problems, but also the scenarios producing these critical situations. Thus, it can be used not only to simplify the study of new investments (reinforcement of existing lines), but also to facilitate the evaluation of hedging strategies and the design of proactive policies to avoid the detected congestion.**

***Index Terms*-- Transmission planning, congestion management, uncertainty, data mining, artificial intelligence techniques, automatic learning, decision trees.**

I. Introduction

Congestion in a power system occurs when one line reaches its maximum transmission capacity. This means a limitation on the transfer capability between several locations on the network, which usually implies a generation redispatch in order to avoid load curtailment. Transmission planning aims at making decisions to increase the transmission capacity of the system in an optimal way so as to prevent the system to experience undesired situations, and in particular to avoid possible congestion problems in the future. As building new lines is almost impossible in most developed countries, transmission system investments focus on reinforcement of existing lines.

Traditionally, utilities have planned, designed and operated their power systems as a whole. In this integrated approach, transmission planning involves technical and economic assessment of various generation and transmission expansion plans. Usually transmission planning has been formulated as a large scale optimization problem [1][2][3]. Because of the very large size of current power systems, the computational burden of this approach is in practice very high, therefore several artificial intelligence techniques (e.g. simulated annealing [4], genetic algorithms [5], heuristic search [6], decision trees [7]) have been used more recently. In this integrated approach, both generation and transmission options are controlled by the same utility so the number of uncertainties to be taken into account is small. Traditional sources of uncertainty are: demand growth, hydraulic conditions, fuel prices, as well as availability of lines and generators.

However, today, with the full liberalization of the energy market, the future becomes much more uncertain. Now, in order to favor competition, the allocation of generation is a matter of independent decision making mainly based on business opportunities. Thus, the number of uncertainties that should be taken into account in planning studies has increased drastically, with new factors such as location and size of new generators, retirement of old ones, or generation patterns. Therefore, within this new environment, new planning approaches, methods and tools are required to cope with these new uncertainties [8][9].

In this paper we propose a novel methodology based on the application of Data Mining (DM) techniques to identify congestion problems under both classical and new uncertainties. DM techniques have emerged in the early nineties and they allow extracting meaningful information from data bases consisting of many different pre-analyzed scenarios [10]. DM has been successfully applied to various problems in power systems (e.g. see [11][12]).

The paper is organized as follows. In Section II, the notion of congestion severity is introduced. In Section III the overall identification methodology is described. In Section IV an illustrative example of the proposed approach is provided. A real case study is shown in Section V. Finally, conclusions are pointed out at the end. Notice that we do not provide any details about Data Mining methods (see e.g. [11][10]). Rather, we show how they can be used to solve particular stages of the methodology proposed to identify congestion problems.

E. F. Sánchez-Úbeda, J. Peco, P. Raymont and T. Gómez are with the Instituto de Investigación Técnológica (IIT), Universidad Pontificia Comillas, Alberto Aguilera 23, 28015 Madrid – SPAIN (e-mail: {Eugenio.Sanchez, Jesus.Peco, Paviel, Tomas.Gomez}@iit.upco.es).

S. Bañales and A. L. Hernández are with the Power System Department Research & Development Division, Electricité de France (EDF), Avenue du Général de Gaulle, 92141 Clamart Cedex – FRANCE (e-mail: banales@alum.mit.edu, Anne-Laure.Hernandez@edf.fr).

II. Notion Of Congestion Severity

Two key issues arise when studying a particular scenario: (i) Is there a congestion?, and (ii) If so, how important is it?. To answer both questions, we propose appropriate congestion indicators as well as two types of numerical simulations.

A. Congestion indicators

The congestion indicators can be used to identify congestion problems in a particular scenario as well as to assess their severity.

Two types of indicators can be used: local and global congestion indicators. The first ones provide information about the congestion of a particular line. However, sometimes an index of the global congestion of the network can be useful. The global congestion indicators summarize the local congestion problems into a more general index. Using these global indicators a network situation can be directly classified according to "its severity" in terms of congestion problems.

Some of the proposed local congestion indicators are: Power flow/ transmission capacity Ratio (F/C), OPF dual variable associated with the transmission capacity of a line, non supplied energy of a bus, etc.

Concerning global congestion indicators, we propose: Mean squared power flow / transmission capacity Ratio, difference between total dispatch costs with and without power flows constraints (redispatch costs, RC), Number of redispatched generators, Total redispatched generation, etc. In the illustrative example we have used the RC indicator because it provides directly an economic assessment of the congestion problems.

B. Types of numerical simulations

At first, a DC non-linear optimal dispatch without constraints on power flows will provide a first diagnosis of the situation, telling us if there are overloaded lines or not. Unfortunately, the results of this first simulation do not assess the severity of the congestion accurately. For example, a scenario with a heavily overloaded line could be less severe than other scenario with other lines being only slightly overloaded because, in the first case, the economic impact associated with the generation redispatch needed to solve the congestion could be much smaller than in the second situation.

A second analysis, consisting in a DC non-linear optimal dispatch with transmission constraints will allow determining the congestion severity in terms of its economic impact more accurately.

Notice that in transmission planning studies, several scenarios have to be considered. For example, when studying a new investment, the resultant network should perform successfully under every plausible scenario. Thus, the severity of a particular congestion depends not only on the analysis of an scenario but also on its importance. This importance is related to both its probability and its impact on the network.

III. Identification Methodology

The proposed identification methodology consists of two main steps (Fig. 1): (i) data base generation of possible scenarios, and (ii) data base analysis using data mining techniques. This approach can be repeated several times as our knowledge about the congestion problem increases.

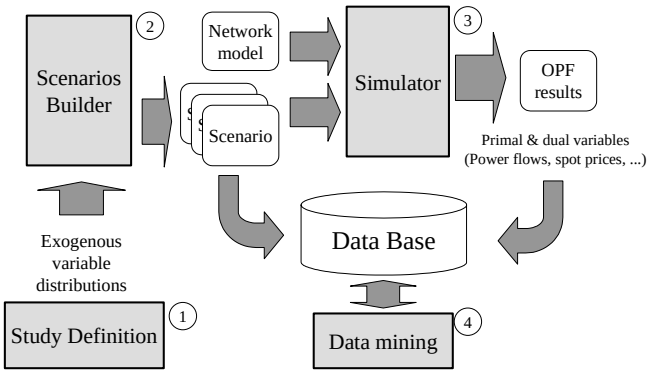

Fig. 1. Identification methodology

A. Definition of the range of possible situations

Uncertainties will be modeled as probability distributions. These distributions can be discrete or continuous, depending on the nature of the variable. The different network situations (scenarios) will be generated by sampling these distributions. This probabilistic model is flexible enough to reflect the intrinsic probabilistic nature of the uncertainties to be modeled (availability of lines and generators, hydraulic conditions, generation costs, location and size of new generators, retirement of old ones, etc). For example, for a given bus one can set a distribution for a given technology describing the probability of having zero, one or two units.

This probabilistic model should incorporate all prior knowledge as well as the definition of the range of conditions which the study aims to cover. As the size of the network increases, both the number of possible scenarios and the computational requirements to simulate them grow exponentially. Thus, in order to deal with this curse-of-dimensionality, the probability distributions of rare events with a significant impact (e.g. the failure of a nuclear unit) are biased to ensure that they will be sampled.

B. Generation and simulation of scenarios

The objective of this stage is to obtain a sufficiently rich data base (DB), which both contains plausible operating states and covers all relevant congested situations. These scenarios are randomly drawn from previous probability distributions. According to previous section, each scenario is simulated using a DC non-linear optimal dispatch with and without transmission constraints. All in all, from each scenario we save in the DB a set of parameters or attributes characterizing it. These attributes are mainly installed capacities of generators (by type and bus) and generation costs, together with congestion indicators.

C. Analysis using Data Mining Techniques

Traditionally, utilities have planned, designed and operated their power systems as a whole. In this integrated approach, sources of uncertainty were small and well known. Typically, with 20 scenarios all relevant situations were captured and the planer knew how to run the network under these circumstances.

However, today, with the full liberalization of the energy market, the future becomes much more uncertain as both the allocation of generation is a matter of independent decision making and bilateral contracts between agents from different countries may change the traditional power flow patterns. Under this new situation, the number of scenarios is huge and previous knowledge about the network behavior is not enough to identify the possible congestion problems, their severity and their root causes.

Data mining techniques allow dealing with this new situation (see e.g. [10][13]). In particular, these techniques are able to identify the possible congestion problems, their severity and their root causes by analyzing huge amounts of simulated scenarios in a systematic way. We have used a set of techniques, including decision and regression trees [11][10], clustering techniques and ORTHO models (e.g. see [14]).

The proposed approach is depicted in Fig. 2. After defining the network model and the considered uncertainties, a large set of scenarios are generated and analyzed using data mining techniques.

These techniques rely on the selection of suitable candidate attributes that are able to explain the root causes of network congestion. This topic is especially relevant when dealing with real sized networks, where it is necessary to identify clusters representing groups of buses (i.e. regions) as well as groups of lines connecting those regions (i.e. transmission paths) by finding similarities among physical parameters (see e.g. [11][15][16][17]). This step is needed to define global variables related to transmission paths and regions (e.g. the generation capacity in a region or the number of gas units in a region) instead of particular lines and buses. Note that these regions are a key issue as excesses or deficits in the generation capacity of *different areas* cause most congestion scenarios.

The next step in our methodology is to perform a global approach to the congestion problem in order to answer two questions: (i) what scenarios, among all possible ones, present congestion problems and why?, and (ii) what are the congestion modes (i.e. congestion patterns and network behaviors) and what are the causes of these different modes?.

First, one can get an idea of the possible congestion problems by building a supervised model such as a decision tree or an ORTHO model to estimate some global congestion indicators.

Second, clustering techniques are used to identify congestion modes (i.e. congestion patterns and network behaviors), and supervised models to know when these patterns appear.

Finally, one can focus on particular congestion modes, congested transmission paths or congested lines by building supervised model to estimate some local congestion indicators.

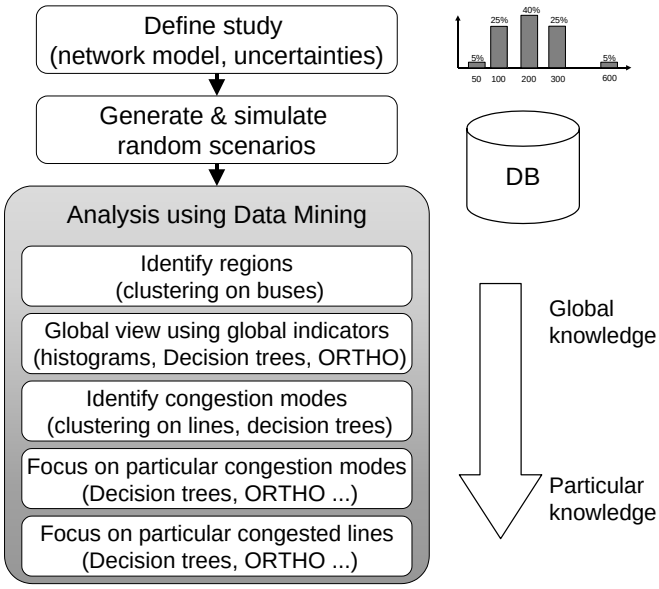

Fig. 2. Identification methodology (detail)

IV. Illustrative Example

In order to illustrate the proposed methodology, we use an academic network consisting of 11 buses and 14 lines (see Fig. 3). In each bus the demand is constant (shown in percentage of the total one), whereas the number of units of each technology can vary (figure shows the maximum installed capacity by technology, where: C is coal, H is hydro, N is nuclear and G is gas). Notice that the number of units could be related to new generators, retirement of old ones, or unit availability. The generation costs of nuclear, gas and coal are drawn from normal distributions N(0.3, 0.025), N(0.55, 0.1) and N(0.6, 0.05), respectively. Finally, two types of hydraulic conditions (dry and humid) have been taken into account.

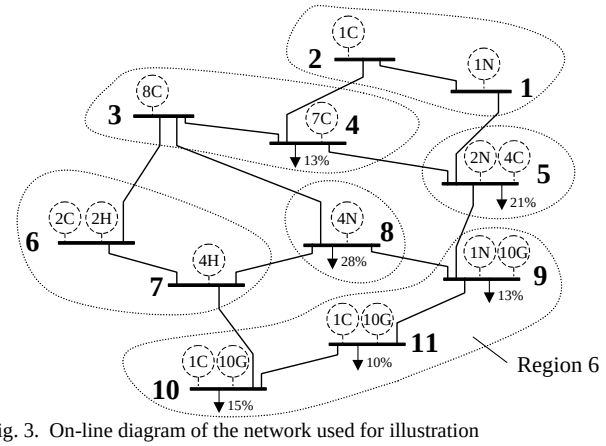

Fig. 3. On-line diagram of the network used for illustration

A large DB consisting of 40000 different scenarios have

been built, where each scenario represents a different network situation generated by random sampling from the probability distributions used to model the uncertainties. Each scenario is described by 67 attributes, corresponding to parameters describing the electrical state (installed capacities per unit and technology, generation costs, number of units by technology and by bus, etc). Using clustering techniques, six regions have been identified (Fig. 3), e.g. region 6 consists of buses 9, 10 and 11. Note that the total installed capacity of each region has been also used as an attribute characterizing each scenario.

In order to obtain a first overview, we have built a decision tree (Fig. 4) to estimate the value of the global congestion indicator RC. We consider the network has no congestion problems if RC is zero (label 'RC=0'), and it has congestion problems (label 'RC>0') otherwise. For example, this decision tree tells us that if we have enough installed capacity in region 6 (ICReg6>34.5), then the network will be clearly congested if coal is cheaper than gas (costC-costG<-0.03), see node A in Fig. 4. To evaluate the generalization capability of these decision trees, they were tested on the basis of an independent test set of 8000 scenarios (i.e. not used for building the trees), yielding an overall error rate of 14.3% (i.e. 85.7% of correct classifications)

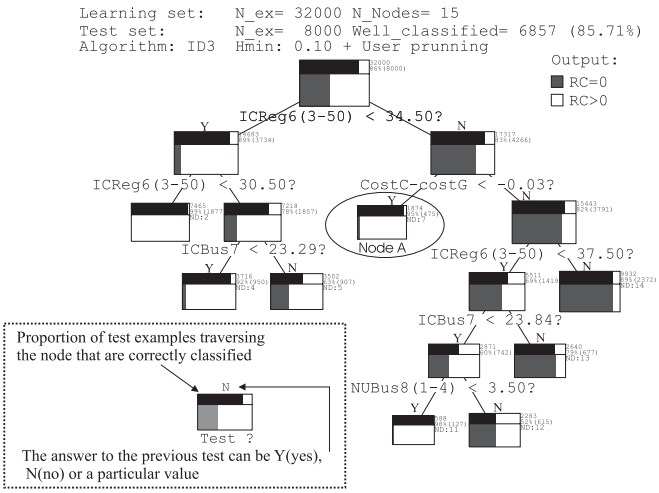

Fig. 4. Decision tree for the global congestion indicator RC

After extracting global knowledge about the congestion problems of the network, we can study separately the congestion of each line. Lines 8-9, 5-9, 3-8, 7-10 and 9-11 have congestion problems. For example, Fig. 5 shows a decision tree to estimate if line 3-8 is congested or not. We consider it is non-congested if the local congestion indicator F/C of line 3-8 is smaller or equal than 1 (N), congested (C) otherwise. For example, this tree tell us that line 3-8 is congested when the number of nuclear units of bus 8 is not maximum (NUBus8<3.5), the scenario is dry (ICBus8<30) and the total number of nuclear units is less than 6.5. This tree yields an overall error rate of 4.24% in the same test set (i.e. 95.76% of correct classifications). Note that better decision trees, in terms of accuracy, can be obtained by allowing the expansion of some nodes.

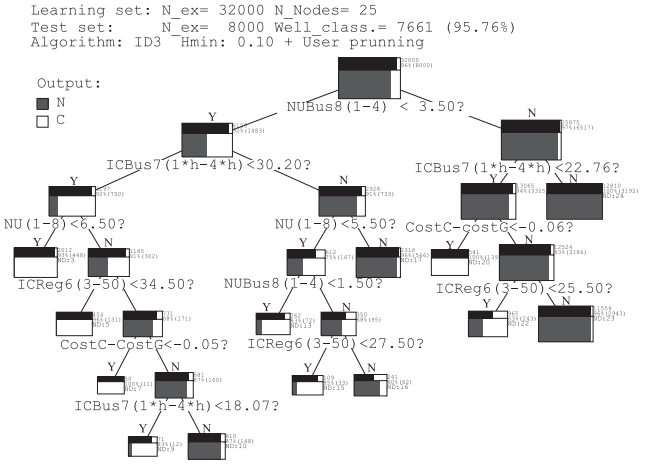

Fig. 5. Decision tree for the congestion of line 3-8

Furthermore, it is possible to estimate the exact value of F/C of line 3-8 by means of the ORTHO model (Fig. 4). This model provides us the sensitivity of the output with the different relevant input variables. For example, this model tell us that if the installed capacity of region 6 is large enough (ICReg6>47) then we can not decrease the power flow of line 3-8 by increasing the installed capacity of this region, which is physically sound. According to this model, higher hydraulic capacity in bus 7 (ICBus7) leads to lower power flows in line 3-8. On the other hand, this flow decreases when the availability of nuclear units in bus 8 (NUBus8) increases.

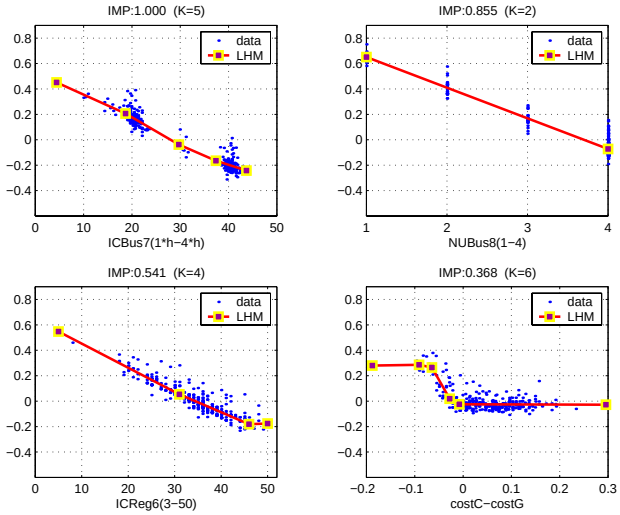

Fig. 6. ORTHO model for the local indicator F/C of line 3-8

V. Real Study

The proposed methodology has also been tested on a realistic problem. In particular we have used a detailed model of the real Spanish network consisting of 1140 nodes, 1500 lines and 225 installed generators. It also includes interconnections with France, Portugal and Morocco. There are 5 different technologies: Coal, hydro, nuclear, and gas.

Several uncertainties has been taken into account: the unavailability of 5 main nuclear units, the possibility of 19

new gas plants comprising a number of units drawn from the uniform distribution U(0, 3); different generation costs of coal and gas technologies, see Fig. 7. Finally, two types of hydraulic conditions (dry and humid) have been taken into account.

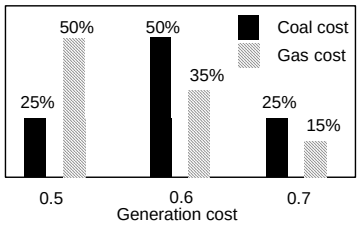

Fig. 7. Distribution for the generation costs of coal and gas technologies

A DB consisting of 6000 scenarios have been built by random sampling from previous probability distributions. In this case each scenario is described by 146 attributes, corresponding to parameters describing the electrical state (see Appendix). Note that this set of attributes also includes more general parameters summarizing the information of different regions (in this study 20 independent regions have been identified).

The obtained models are simple, accurate and physically sound, providing physical understanding of the particular congestion problem. For example, Fig. 8 shows a decision tree to estimate if a line L is congested or not. We consider it is non-congested if the local congestion indicator F/C of line L is smaller or equal than 1 (N), congested (C) otherwise. According to this model, this line is congested in the next three cases (Fig. 9): (i) when there are no gas units in Bus 1082 (GUBus1082<0.5) and it is a dry scenario; (ii) when there are no gas units in Bus 1082, the scenario is humid (Hcoef<0.92) and there is not enough installed capacity in region 11, and; (iii) when there are no gas units in Bus 1082, the scenario is humid, there is enough installed capacity in region 11 (ICReg11<17.65), but there are no gas units in Bus 775 (GUBus775<0.5).

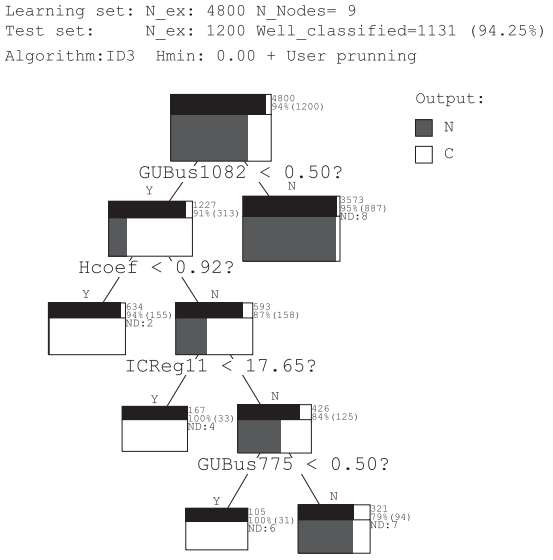

Fig. 8. Decision tree for the congestion of line L (real case study)

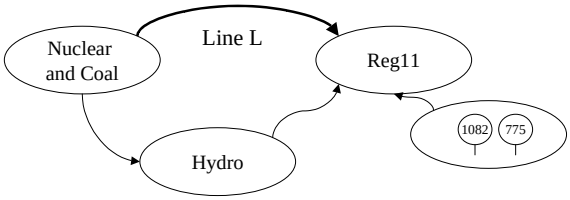

Fig. 9. On-line diagram of the real case network (simplified)

VI. Conclusions

In this paper we have proposed a methodology to identify congestion problems under both "traditional" and "new" uncertainties such as uncertainty about generation costs, location and size of new generators, retirement of old ones, generation patterns, etc. The methodology allows not only identifying the transmission paths and corridors which will have congestion problems, but also the scenarios producing these critical situations. Thus, it can be used not only to simplify the study of new investments (reinforcement of existing lines), but also to facilitate the evaluation of hedging strategies and the design of proactive policies to avoid the detected congestion.

Concerning future work, we are extending the proposed methodology to assess the economic impact of a congestion problem. This last step is required to propose and evaluate alternative investment plans.

VII. Appendix - Candidate Attributes

Here we provide a brief list of possible candidate attributes that has been used in our studies.

- Installed capacities per unit and by technology
- Generation costs by technology
- Number of units by technology and by bus
- The difference of cost between technologies
- The total installed capacity by technology
- The total installed capacity by bus
- The cheapest technology by bus
- The most expensive technology by bus
- The global installed capacity by technology
- The total installed capacity by region
- The total installed capacity by region and by technology
- Hydraulic conditions

VIII. References

[1] L. 1. Garver, R. Villasana, S. J. Salon, "Transmission Network Planning Using Linear Programming," *IEEE Trans. on PAS*, vol 104, No. 2, Feb. 1985.

[2] V. Levi, M. S. Calovic, "A New Decomposition Based Method for Optimal Expansion Planning of Large Transmission Networks," *IEEE Trans. on Power Systems*, vol 6, No. 3, August 1991.

[3] R. Romero, A. Monticelli, "A Hierarchical Decomposition Approach for Transmission Network Expansion Planning," *IEEE Trans. on Power Systems*, vol 9, No. 1, Feb. 1994.

[4] R. Romero, R. A. Gallego, A. Monticelli, "Transmission System Expansion Planning by Simmulated Annealing," *IEEE Trans. on Power Systems*, vol 11, No. 1, Feb. 1996.

[5] H. Rudnick, R. Palma et al., “Economically Adapted Transmission Systems in Open Access Schemes- Application of Genetic Algorithms,” *IEEE Trans. on Power Systems*, vol 11, No. 3, Aug. 1996.
[6] G. Latorre-Bayona, I. Pérez-Arriaga, “Chopin, a Heuristic Model for Long Term Transmission Expansion Planning,” *IEEE Trans. on Power Systems*, vol 13, No. 2, Feb. 1998.
[7] J. Peco, E. F. Sánchez-Úbeda, T. Gómez, “Enhancing optimal transmission and subtransmission planning by using decision trees,” BPT99-304-16, in *Proc*. IEEE Power Tech, August, 1999.
[8] CIGRE TF 28.05.05, “Techniques for power system planning under uncertainties,” Cigré report 154, April 2000.
[9] T. de la Torre, J. W. Feltes, T. Gómez, H. Merrill, “Deregulation, Privatization, and Competition: Transmission Planning under Uncertainty,” *IEEE Trans. on Power Systems*, vol 14, No. 2, May 1999.
[10] E. F. Sánchez-Úbeda, “Models for data analysis: contributions to automatic learning,” Ph.D. dissertation, Univ. Pontificia Comillas, 1999.
[11] L. Wehenkel, *Automatic Learning Techniques in Power Systems*, Kluwer Academic, Boston, 1997.
[12] L. Wehenkel, "Machine learning for power system security assessment," IEEE Expert, Intelligent Systems and their Applications, vol 12, No. 5, 1997, pp. 60-72.
[13] V. Cherkassky and F. Mulier, *Learning from data: concepts, theory, and methods*, John Wiley & Sons, New York, 1998.
[14] E. F. Sánchez-Úbeda and L. Wehenkel, "Automatic fuzzy-rules induction by using the ORTHO model," in *Proc.* IPMU'2000, July 2000.
[15] T. Kohonen, *Self-Organizing Maps*, Springer Verlag Series in Information Sciences, vol. 30, 1995.
[16] R. O. Duda and P. E. Hart, *Pattern classification and scene analysis*, John Wiley and Sons, 1973.
[17] L. Kaufman and P. J. Rousseuw, *Finding Groups in Data*, John Wiley & Sons, 1990.

IX. BIOGRAPHIES

Eugenio Francisco Sánchez Úbeda received the Electronic Engineering Degree in 1991 and the Ph.D. Degree in 1999, both from the Universidad Pontificia Comillas. He is presently a research fellow and co-ordinator of the Intelligent Systems Group of the ‘Instituto de Investigación Tecnológica’ and teaches a course on “Advanced data analysis” and on “Sorting and searching techniques”. Since 1991, he has been working in the field of artificial intelligence and its applications.

Jesús Pascual Peco González was born in Las Palmas, Spain, in 1972. He obtained his Electrical Engineering Degree in the Pontificia Comillas University of Madrid in 1996. He joined the ‘Instituto de Investigación Tecnológica’ at the Pontificia Comillas University in 1996, where he is currently finishing his PhD thesis on distribution planning models. He has carried out several researches in the areas of distribution planning, distribution regulation, reliability and optimization techniques.

Tomás Gómez San Román obtained his Electrical Engineering Degree in the Pontificia Comillas University, Madrid, in 1982, and his Ph.D. Degree in the Polytechnic University of Madrid, in 1989. He is a research fellow, and has been the director of the ‘Instituto de Investigación Tecnológica’ since 1994. He has been involved in more than thirty research projects with Spanish and European utilities. His areas of interest are: the planning and operation of transmission and distribution systems, power quality and regulatory issues.

Paviel Raymont obtained the Civil Engineering degree from the "Ecole des Mines de Nancy" (France) in 1999. He is currently working for Electricité de France (EDF), and sent in the Instituto de Investigación Tecnológica (IIT), to work on the project concerning network planning with uncertainty in a deregulated environment. His areas of interest are related to numerical simulation, power systems and management of industrial systems.

Santiago Bañales López holds a Master of Science (1998) in Technology and Policy from the Massachusetts Institute of Technology and a dual Engineering degree from the Escuela Técnica superior de Ingenieros Industriales de Madrid and from Ecole Centrale Paris (1994). He has been Researcher at the Power Systems Department of the Electricité de France Research and Development Division until December 2000, where he specialized on transmission planning and pricing issues under open access. Currently he is part of the Utilities practice of A.T. Kearney, a management consulting firm.

Anne-Laure Hernández graduated from the “Ecole Nationale Supérieure des Mines de Paris” (ENSMP) in 1993. She joined the R&D division of EDF in 1994, where she first worked with the Electricity Uses Department. Since 1998, she works with the Power System Department at EDF-R&D, where she is mainly involved in Transmission System development.