Proceedings of the 5th WSEAS Int. Conf. on SIMULATION, MODELING AND OPTIMIZATION, Corfu, Greece, August 17-19, 2005 (pp243-249)

DATA MINING AND SIMULATION PROCESSES AS USEFUL TOOLS FOR INDUSTRIAL PROCESSES

ORDIERES MERÉ, J.(1); ALBA ELÍAS, F.(1); GONZÁLEZ MARCOS, A.(2); CASTEJÓN LIMAS, M.(2); MARTÍNEZ DE PISÓN ASCACÍBAR, F.J.(1)
(1) Department of Mechanical Engineering. University of La Rioja. c/Luis de Ulloa s/n, 26004 – Logroño (LA RIOJA). SPAIN
(2) Department of Electrical Engineering. University of León. Campus de Vegazana s/n, 24071– León (CASTILLA Y LEÓN). SPAIN

Abstract: - The most common goal of the factory owner is to achieve better quality in the final product by means of an improved process control. The significance and relevance of optimizing the existing control models is even greater in the open-loop control systems or in those governed by computational methods dependent on adjustable parameters. This talk reviews some typical industrial environments and focuses on some parts of them in order to show the real interest of these improvements. We will identify some difficulties in obtaining these improvements and show how the optimal control model for the manufacturing process can be obtained from data provided by sensors. We will also discuss some technical problems that are related to the main goal, and will identify some topics concerning outliers, density and topology. Also, we will show how these techniques can be applied as an instrumental toolbox in addressing some environmental problems.

Key-Words: - industrial applications; neural networks; outliers; density; improvement process control; advanced quality control.

1 Introduction

When the people think about industrial processes, from a supervision point of view, the idea of automatic control with microcontrollers and PLC systems arise as the right way. However, there are some industrial processes where this classical approach doesn't work.

When the process of producing hot steel coils is considered (Fig. 1), it is usual to get speeds about 15m/s in the last steps of the process. It means 15mm each millisecond, so if the objective is to produce a controlled width, it becomes necessary to measure, to compute and to order to the hydraulic systems to move the cylinders all under 0,1 millisecond. Currently this is not possible, so another control strategy is required instead those based on closed control loop.

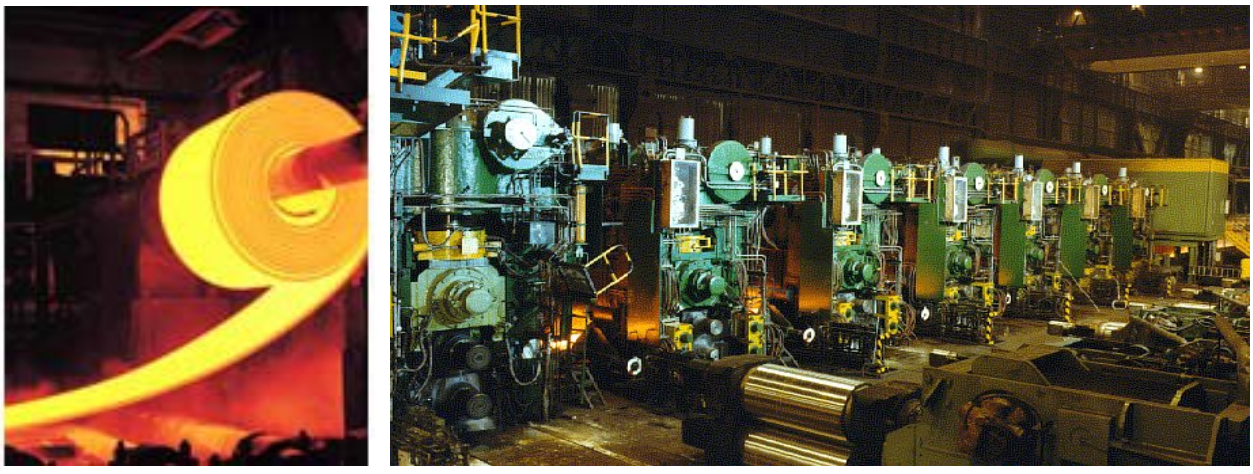


Fig. 1. Different sections of a steel line.

In this case an open loop control strategy is mandatory and then a model must predict a set of consigns to be established at low level controllers and after the new coil is produced an estimation of error is carried out and the model will be informed for a sharp estimation of setups.

Other type of processes where a direct control strategy is not suitable, are those where the goal is to build a product with specified parameters which are not directly measurable. The measuring process is carried out offline in a laboratory and downwards the process. Also in these cases an open control strategy must be adopted.

Just as an example, the process of mixing components for producing rubber (Fig. 2) for profiles of automotive industry (Fig. 3) is inside this category:

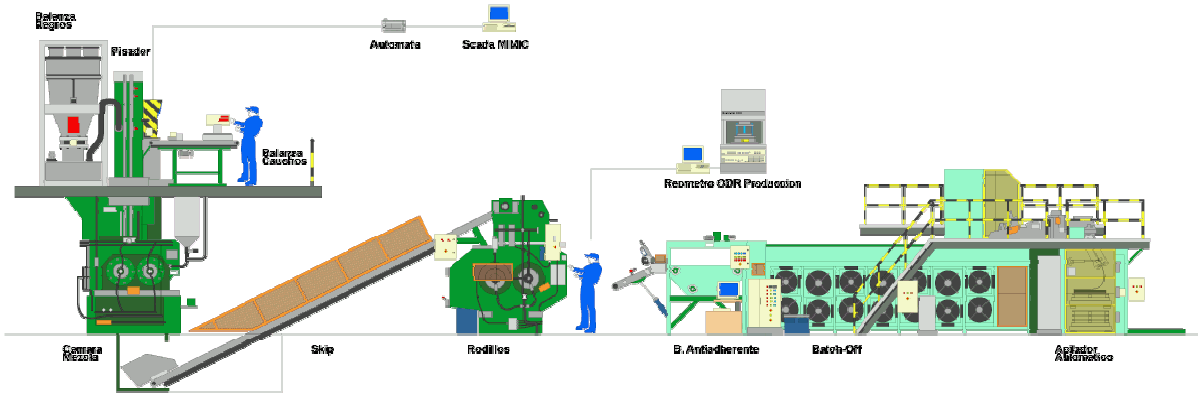
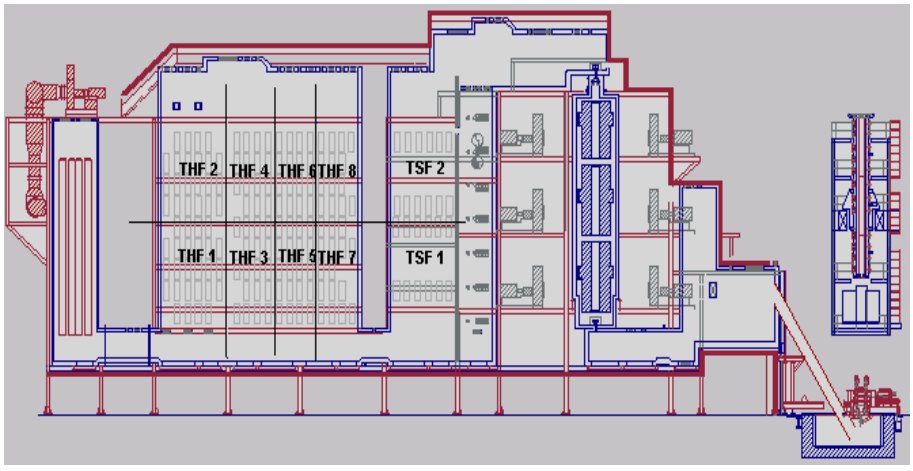


Fig. 2. Rubber mixing process.

Fig 3. Rubber profiles for automotive industry.

On this type of processes the goal is to produce a rubber with a homogeneous structure and with a viscosity predefined, but, unfortunately the viscosity is measured off line, at laboratory.

These types of processes are quite usual, for example mechanical properties of hot dip galvanizing coils are measured after coating the coil, by using a destructive method. In this case a system to estimate these mechanical properties taking into account coil composition, thermal cycle inside the furnace, speed, etc. will be very useful.



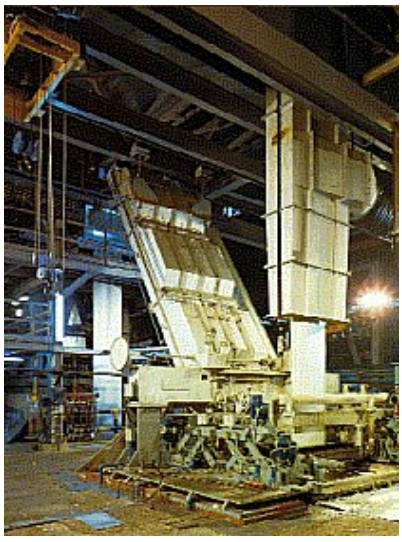


Fig 4. Annealing furnace scheme and molten zinc bath section from a galvanizing line.

When the processes are running, Data Mining (DM) can be seen as a strategy, coming from Knowledge Discovery arena, to be explored.

The main part of data mining is concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. The idea is that it is possible to strike gold in unexpected places as the data mining software extracts patterns not previously discernible or so obvious that no-one has noticed them before. The analysis process starts with a set of data, uses a methodology to develop an optimal representation of the structure of the data during which time knowledge is acquired. Once knowledge has been acquired this can be extended to larger sets of data working on the assumption that the larger data set has a structure similar to the sample data. This is analogous to a mining operation where large amounts of low grade materials are sifted through in order to find something of value.

Data mining is not a product that can be bought. Data mining is a discipline and process that must be mastered - a whole problem solving cycle.

If it is quite clear that data warehousing provides the enterprise with a memory, it could be stated that **Data mining provides the enterprise with intelligence.**

Just as an example, we can think about one plant producing galvanized coils for automotive industry. This is a type of industry very adjusted also in tooling involved in manufacturing process.

When a coil, by error, is processed using a harder material than normal and ends up with a client, significant damage may be produced in its factory tooling as processing this harder coil will require higher pressure and bigger forces are involved. So, presses can become broken and other problems arise.

It is necessary to avoid these errors. An “artificial lock” needs to be provided in order to ‘predict’ the elongation regarding the tension and pressure used in the skin-pass. If the predicted elongation is a lot different than the measured one, the coil must be then removed from the queue for further analyses. This is a typical application of DM as far as a model from processed coils needs to be derived before to be used on this “artificial lock”.

The methodology used on that application is shown graphically:

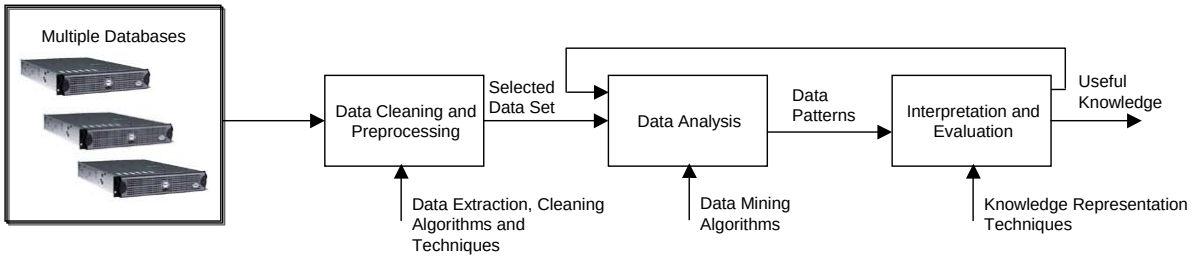


Fig 5. Data mining methodology.

Obviously this paper doesn’t support the idea of “Data Mining everywhere and every time”. If an application needs a model that can be provided by classical tools it will be preferred, as far as this procedure is less energy consuming than those linked to DM methodologies.

2 Opportunities

In despite of classical methodologies like CRISP-DM, CRITIKAL, SESAME, etc., and also their neutral capabilities regarding specific tools for data processing, there seems to be a direct relationship between its potential benefits and the quantity of often-contradictory claims, or myths, about its capabilities and weaknesses.

Data mining can lead to significant change in several ways. First, it may give the talented manager a small advantage each year, on each project, with each customer/facility. Compounded over a period of time, these small advantages turn into a large competitive edge.

Experience in building models, however, can ensure more profitable use of data mining, since data mining is simply the newest tool for building models.

The less domain knowledge a data mining expert brings to a problem, the more important it is to perform the data mining in close cooperation with people who understand the business. This is why normally in our projects deeply cooperation is carried out between ‘data – crackers’ and

technological people coming from the process itself trying to setup a powerful team.
Tools used can not be the same from project to project, taking into account specific goals, type of data available, type of knowledge to be discovered and way for implementation of this knowledge.
Data mining is most cost-effective when used to solve a particular problem. Although a data- mining tool can indeed explore your data and uncover relationships, it still needs to be directed toward a specific goal. *Simply giving a data-mining tool a mailing list and expecting it to find their profiles that improve the expectation of next business is not particularly effective.*
Data mining is useful wherever data can be collected. Of course, in some instances, cost/benefit calculations might show that the time and effort of the analysis is not worth the likely return.
The algorithms of data mining are complex, but new tools have made those algorithms easier to apply. Often, just the correct application of relatively simple analyses, graphs, and tables can reveal a great deal about our problem. Much of the difficulty in applying data mining comes from the same data-organization issues that arise when using any modeling techniques. These include data preparation tasks--such as deciding which variables to include and how to encode them--and deciding how to interpret and take advantage of the results.
Another problem is try to discover new relationship among several variables where there is not. This is a tricky but usual problem.
More data items are useful only if they contribute more information about the issues at hand, or goals. Otherwise, they can be worse than worthless. A database may have a great deal of information about an item (or about the relationship between items) but nothing about other items that are actually closely related.
Even when building a massive database, it helps to try out some simple analysis on the data while the database is still moderate in size. After the analysis, a decision for collecting the data differently or to collect different data altogether can be taken.
Working on data mining normally means to do predictive modeling, so to have a variable that is being predicted from other variables. Or it means clustering where the goal is just finding groups in the data. Dependency modeling is when we do a density estimate. Basically, it is trying to model the joint probability density that generated the data in the first place, a much harder problem; some techniques work. Another is summarization, which looks for relations between fields or associations. Sometimes finding correlation and pieces in the data can be useful.
Finally, the last class of techniques accounts for sequence. It turns out that there are amazingly efficient algorithms that will do things like find you all frequency concerns in there, which is an interesting reduction. A lot of people are working on trying to relate this back to classical analysis techniques. There are a lot of interesting things that can be done when the goal is to account for the sequence in data and changes in data.

3 Main Concerns

Looking a bit deeply into industrial problems particular aspects are envisaged:
Normally utility companies have archived many years of system performance data, including reliability data. Often, state regulatory agencies require this data to be reported annually. Once done, it may never be referenced again. Yet, if analyzed, valuable trend information could be identified to support operations decisions and justify equipment expenditures.
As electric utilities enter the era of changing regulation, decisions regarding expenditures on new services and maintenance are being revisited. In some situations, the pursuit of higher shareholder returns has resulted in reduced line maintenance budgets, so service reliability has suffered. In response to pressure from customers and Public Utility Commissions, maintenance budgets are being restored to improve outage indices.
The first step to improve reliability or power quality in general, is to apply a consistent method of measurement. In most cases, utilities use reliability indices defined by the IEEE Trial Use Guide for Electric Power Distribution Reliability Indices developed under the working group on system design.
Data should be accumulated consistently for relative comparison purposes. When considering distribution system reliability improvement, distribution engineers face many alternatives. A big choice is to decide how many resources to commit to preventing faults as opposed to responding to them. Budget constraints, operating constraints, overcurrent protection philosophies, and organizational issues may enter into the decision process. Many engineers look to improved maintenance and an overall reduction in the number of faults as the best way to improve reliability. In some cases this is the correct approach.
In other situations, it may be a more costly, less predictable, and slower method of improving

reliability than focusing on improving the overcurrent protection system.
A review of system-wide outage data is the first step in developing a reliability improvement strategy. A breakdown by operating region, by feeder, and if possible, by feeder section over several years is the ideal way to establish averages and trends. The data can then be organized so statistical techniques can be applied to aid in the decision process.
Statistical Process Control (SPC) is a methodology for monitoring a process to identify special causes of variation, and signal the need to take corrective action. If reliable distribution of electric power is viewed as a process, then reliability data can be plotted as control charts. Control charts are used to establish a state of statistical control, monitor a process, and signal when the process goes out of control.

Upper and lower control limits are an important part of SPC analysis. Control limits represent the range between which all points are expected to fall. If any points fall outside the control limits, or if any unusual patterns are observed, some special cause has probably affected the process.
The key point here is to determine these limits taking into account things like technological improvements, environmental conditions, and so on, avoid taking wrong decisions about investments launched by a bad position of limits.
Also these fields provide some classes of problems like those marked as classical: outliers, missing values, duplicate data, inconsistent values and other much more specialized like the “process’ outliers”.
In order to show this particular aspect we can see figure 6 showing control variables from a hot dip galvanizing line.

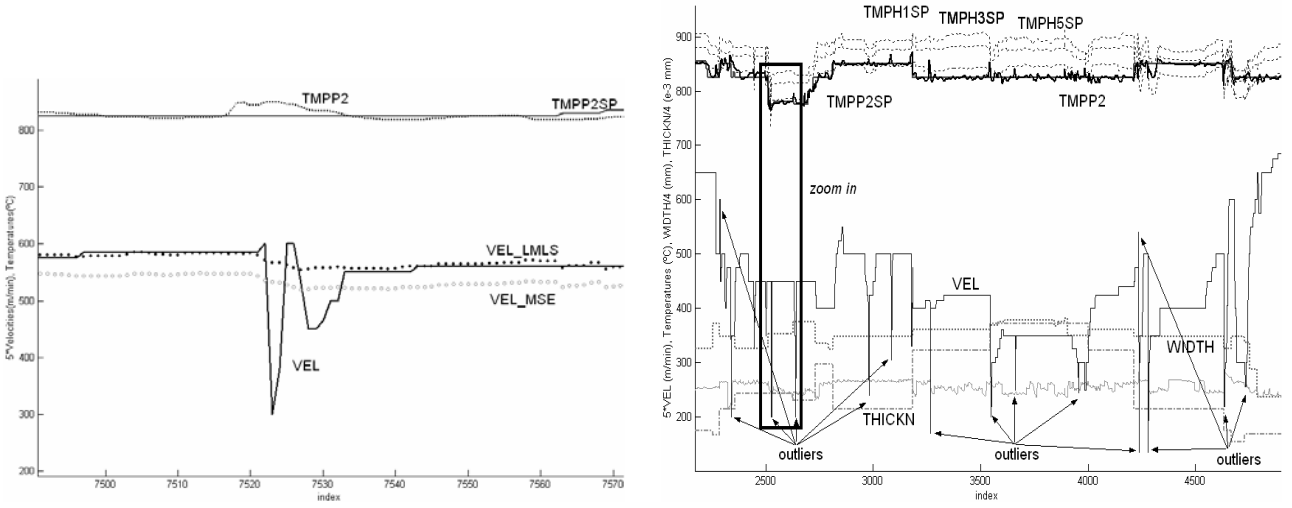


Fig 6. Control variables from a hot dip galvanizing line.

It is common to find ‘measurement errors’ as far as some variables are measured by physical sensors very quickly and in harsh environments (high temperatures, wet conditions, high pressures, etc.), but there are other sources of samples to be managed carefully.
In other cases the process is stopped by other problems, e.g. welding problems during coil extension as shown below, and also in these cases, even when there are no measurement errors, it would be necessary to identify these points in a sample set to be sure that the used strategy doesn’t affect the learning process being carried out. Especially, if they are interfering with data used to build a model, e.g. to estimate line speed when the material format is changing and the temperature needs to be under control.

4 Dealing with outliers

When trying to identify these outliers, a problem arise as far as in those cases where an indirect, automatic control system tries to keep the process under control, the errors are usually non-normal, so outlier identification must be managed carefully.
This means from the scientific point of view that specific algorithms are required for outlier items in a multidimensional space with non normal distribution:

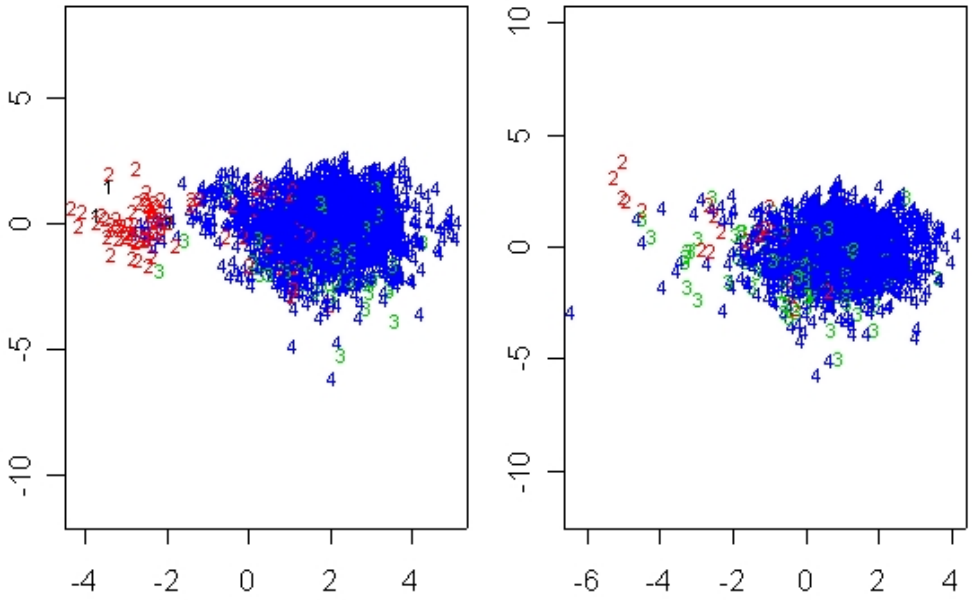


Fig 7. Multidimensional space with non normal distribution.

5 Conclusion

The basic idea is to take part of the pattern recognition algorithms and ask how people do analysis. The traditional method is to take data out of the database. You create your own infrastructure to do analysis. You extract the data, and you start running these scripts and so forth. And soon enough you have created a whole bunch of droppings, and if you come back to this session two weeks later, you don't recall what the files meant, and so forth. That's called a data management problem. It's exactly what a database was created to solve. So the whole idea is to decompose these operations in such a way that a lot of the things can live on the server.

Acknowledges

We will thanks the support from the CEUTIC (Interreg IIIA), the national funded program DPI2004-07264-C02-01, the II Plan Riojano de I+D for their partial funded of this work and for making possible the diffusion of results.

References

[1] Ordieres, J.B., Vergara, E.P., Capuz, R.S., Salazar, R.E. “Neural network prediction model for fine particulate matter (PM2.5) on the US-Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua)”. Environmental Modelling and Software. doi:10.1016/ j.envsoft.2004.03.010.

[2] Espinoza, A.V., Ordieres, J., Martínez de Pisón, F.J., González Marcos, A. “Tao-robust backpropagation learning algorithm” Sent for publishing to “Neural Networks “ journal.

[3] Castejon Limas, M., Ordieres Meré, J.B., Martínez de Pisón, F.J. and Vergara, E. "Outlier Detection and Data Cleaning in Multivariate Non-Normal Samples: The PAELLA Algorithm." Data Mining and Knowledge Discovery, 9, 1--16, 2004.

[4] Ordieres Meré, J.B., González Marcos, A., González, J.A., Lobato Rubio, V. "Estimation of mechanical properties of steel strip in hot dip galvanising lines." Ironmaking and Steelmaking , vol 31 nº 1, 43-50, 2004.

[5] Ordieres, J., López, L. M., Bello, A. et al. ”Intelligent methods helping the design of a manufacturing system for die extrusion rubbers”. International Journal of Computer Integrated Manufacturing. (2003) 16 : 173-180

[6] Ortega, F., Menéndez, C., Ordieres J. and Montequín, V. ”Analysis of Heat Transference in the Regenerative Exchanger of a Thermal Power Plant”. Neural Comput. & Applic. (2000) 9 : 218-226. ISSN: 0941-0643

[7] Ankerst, M., Boeing. “Cooperative data mining: Tightly integrating data mining with visualization”. M2004, the 7th annual Data Mining Technology Conference. Las Vegas. USA.

[8] Duling, D., SAS. “Computational Performance in Data Mining”. M2004, the 7th annual Data Mining Technology Conference. Las Vegas. USA.

[9] Georges, J. SAS. “Using non-numeric data in parametric prediction”. M2004, the 7th annual Data Mining Technology Conference. Las Vegas. USA.

[10] Kahn, B., Capital One. “Why Data Mining is Not Used and Why Better Data Mining Won't Help”. M2004, the 7th annual Data Mining Technology Conference. Las Vegas. USA.

[11] Alhoniemi, E., Hollmn, J., Simula, O., and Vesanto, J. “Process monitoring and modeling using the self-organizing map”. Integrated Computer-Aided Engineering, 1999 6(1): 3-14.

http://www.thearling.com/text/dmtechniques/dmtechniques.htm

http://wwwmaths.anu.edu.au/~johnm/dm/dmpaper.html

http://www.bettermanagement.com/library/library.aspx?libraryid=9175

http://www.cs.helsinki.fi/research/fdk/datamining/

http://www.crisp-dm.org/

http://www.dmg.org/

http://sourceforge.net/projects/pmml

http://www.kcl.ac.uk/neuronet/about/clubs/nl2009.html