Data Mining in industrial processes

Joaquín B. Ordieres Meré,
Fco. Javier Martínez de Pisón Ascacíbar
Grupo EDMANS[1]
Área de Proyectos de Ingeniería
Dpto. de Ing. Mecánica
Universidad de La Rioja
Luis de Ulloa, 20
26004 Logroño (La Rioja)
joaquin.ordieres@dim.unirioja.es,
fjmartin@dim.unirioja.es

Manuel Castejón Limas,
Ana González Marcos
Grupo EDMANS[1]
Área de Proyectos de Ingeniería
Dpto. de Ing. Eléctrica y Electrónica.
Universidad de León. León, 24071.
manuel.castejon@unileon.es, ana.gonzalez@unileon.es

Abstract

The most common goal of the factory owner is to achieve better quality in the final product by means of process improvements. 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 paper 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.
Obviously these techniques can provide with valuable information to the maintenance managers as far as they are capable to adapt specific preventive maintenance strategies to their plants.
We will also discuss the importance of numerical simulations as far as it can be applied both to the process and to the plant in order to produce improvements.

1. Introduction

When people think about industrial processes, from a supervisory point of view, the idea of automatic control with microcontrollers and PLC systems usually seems to be the best method. However there are some industrial processes in which this classical approach doesn't work.
When the process of producing hot steel coils (figure 1) is considered, it is common to register speeds of around 15m/s in the final steps of the process.
That speed is equivalent to 15mm each millisecond, so, if the objective is to produce a controlled width, it becomes necessary to measure, compute and order the hydraulic systems in order to move the cylinders all in less than 0.1 millisecond. Currently this is not possible, so another control strategy is required instead of those based on closed control loops.
In this case, an open loop control strategy is mandatory and later, a model must predict a set of commands to be established on the low level controllers. After the new coil is produced, an estimation of error is carried out and the model is informed for a precise estimation of setups.
Other types of processes where a direct control strategy is not suitable are those where the goal is to build a product whose specified parameters are not directly measurable. The measuring process is carried out off-line in a laboratory. Also, in these cases an open control strategy must be adopted.
Just as an example, the process of mixing components for producing rubber for profiles in the automotive industry falls into this category.

[1] Engineering Data Mining And Numerical Simulations (EDMANS).

Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005, pp.57-66
ISBN: 84-9732-449-8

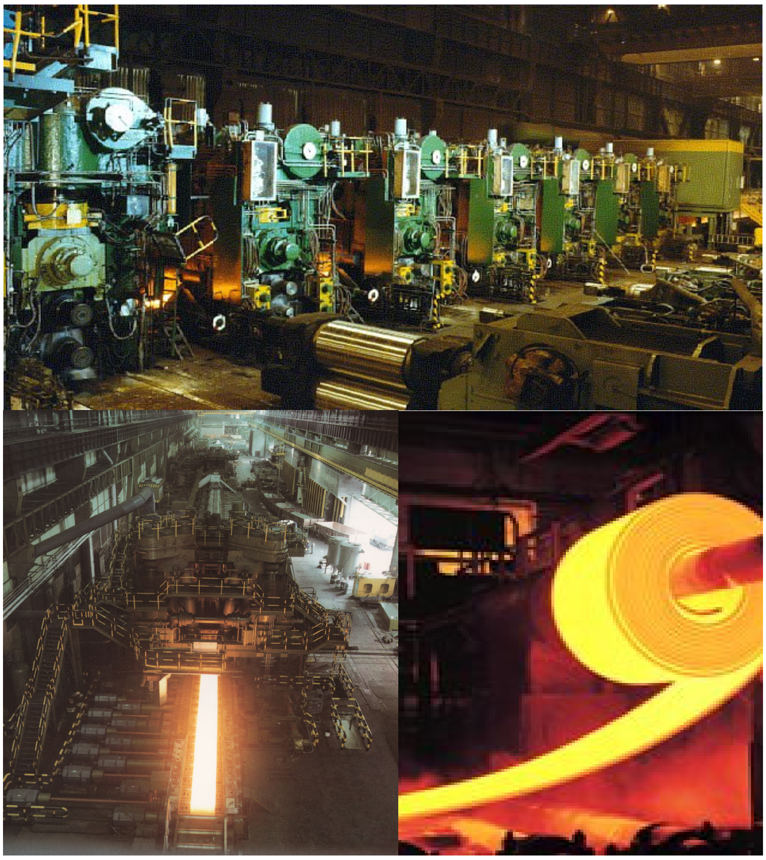

Figure 1. Industrial Steel Processes.

In this case, it is ideal for the system that estimates these mechanical properties to take into account the coil's composition, thermal cycle inside the furnace, speed, etc.

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

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 nobody has noticed them before.

The analysing process starts with a set of data and uses a methodology to develop an optimal representation of the structure of the data during which time useful data is acquired.

Once useful data has been acquired, it can be extended to larger sets of data operating on the assumption that the larger data set has a structure similar to the sample data. This is analogous to a mining operation in which 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; it is a whole-problem solving cycle.

If data warehousing provides the enterprise with a memory, one could say that Data mining provides the enterprise with intelligence.

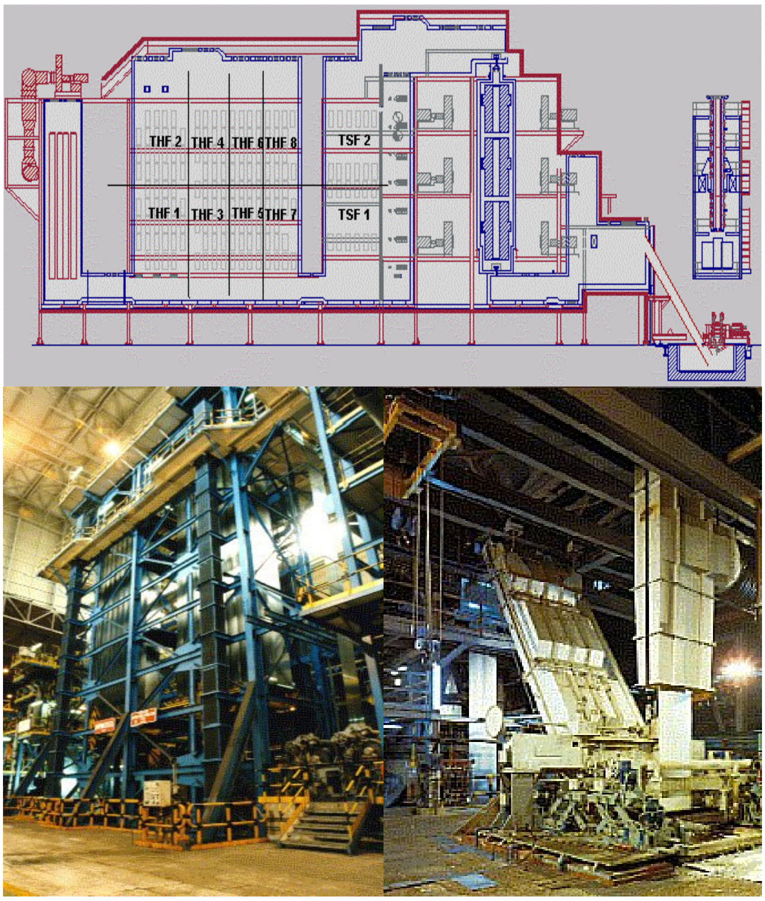

Figure 2. Hot Dip Galvanizing Line (HDGL).

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

When a coil is processed using a harder material than normal by error 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 greater forces are involved. Presses can break, and other problems can 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 quite different than the measured one, the coil must be then removed from the line for further analyses. This is a typical application of DM as far as a model from processed coils needs to be previously derived in order to be used with this “artificial lock”. The methodology used in that application is shown graphically in figure 3.

Obviously, this paper doesn’t recommend the idea of Data Mining wherever and whenever. If an application requires a model that can be provided by classical tools, then that is preferable insofar as this procedure is “less energy consuming” than those linked to DM methodologies.

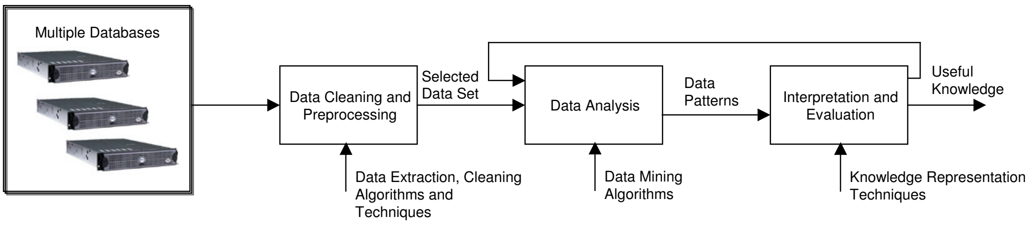

Figure 3. DM Methodology Applied in Industrial Cases.

2. Opportunities

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

Data mining can yield significant, positive changes 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. For that reason, the “data–crackers” and technology experts normally work closely together on our projects as one team.

The 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 obtained and ways to implement that 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 reveal 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 profiles that improve the expectation of the 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 modelling technique. These include data preparation tasks--such as deciding which variables to include and how to encode them and deciding how to interpret and make use of the results.

Another problem is to try to discover new relationship among several variables where relationship doesn’t exist. This is both a tricky and frequent problem.

More data items are useful only if they contribute more information to the issues at hand. If irrelevant data is introduced in the process, it can be worse than worthless. A database may have a great deal of information concerning 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 analyses 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 made.

Working on data mining normally means to perform predictive modelling, or to have a variable that is being predicted from other variables. It can also mean clustering, when the goal is just finding groups in the data. Dependency modelling is when we do a density estimate. Basically, it is trying to model the joint probability density that the data generated in the first place, a much harder problem where only some techniques work. Another is summarization, which looks for relationships between fields or associations. Sometimes finding correlations and pieces in the data can be useful.

Finally, the last class of techniques accounts for the sequence. It turns out that there are amazingly efficient algorithms that will do things like find you all the frequencies concerned, 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. Improvement of Classical Strategies

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.

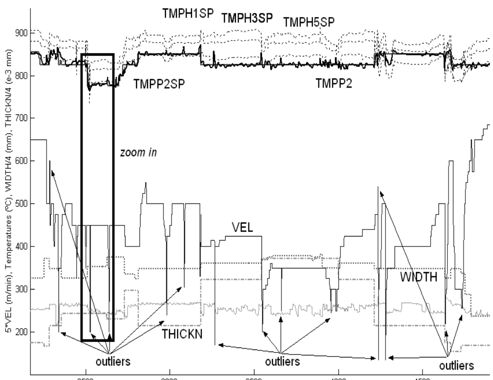

Figure 4. Control Parameters in a HDGL.

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.

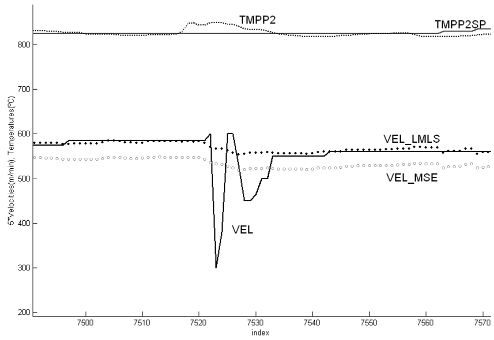

Figure 5. Temp and Velocity in a HDGL.

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 figures 4 and 5 showing control variables from a hot dip galvanizing line.

It is common to find "measurement errors" since 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. Also, in these cases, even when there are no measurement errors, it is necessary to identify these points in a sample set to be sure that the applied strategy doesn't affect the learning process being carried out. It becomes especially important to identify these points 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 arises in those cases where an indirect, automatic control system tries to keep the process under control. The errors are usually abnormal, so outlier identification must be managed carefully.
This means that from a scientific point of view the specific algorithms are required for outlier items in a multidimensional space with abnormal distribution.
The presence of outliers in a data set causes immediately a worse fit, sometimes far from the optimal one, and thus many researchers [17] for an overview) have focused on the detection of these "outliers" that do not follow the pattern of most of the data [8].
As the pattern is latent, it must be estimated from the data set, and thus outliers are involved in the calculation of the general pattern. This obstacle hides the presence of outliers in two different ways, namely masking and swamping [14], turning the task of obtaining a correct approximation of the structure into a really difficult one.
These two effects are related to distortions caused in the location estimator and the shape of the metric used in the analysis, the most common one being the Mahalanobis metric.

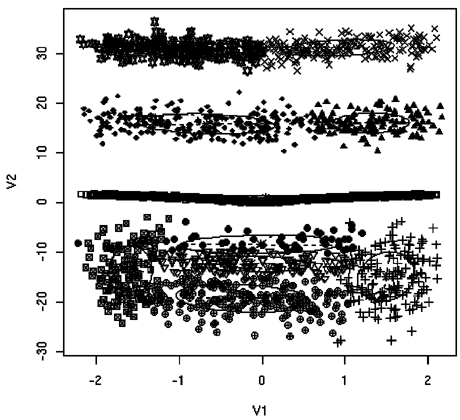

Figure 6. Identifying a structure among outliers.

In order to deal with these limiting aspects we were producing a new algorithm [4] trying to fill the gaps in the available algorithms where data sets do not follow a Gaussian distribution and no a priori metric can be assumed to set up different models.
We feel compelled to reject any dependency on any a priori metric because most of the times the analyst does not have any evidence of such "correct" metric and the results must be similar, irrespectively of the unit system of the samples or the linear transformations the data set might have suffered.
This frequently forces the analyst to affine equivariant estimators of location and shape [16] [14].
In some cases it is possible to identify a structure among outliers which make much more difficult to isolate these elements as we can see in figure 6.
Within the process, some shifts from the sensors can make such situations feasible, as the outlier tools identify several sub-models instead of several groups of outliers. We were working on introducing modellers based on neural network technology, using robust concept as involved in weight estimation [6].
This is just an example of the main topic presented here which is the opportunity to make techniques work together against a problem instead of seeing them as conflicting.
Obviously, this approach carries out problems related to operational strategies usually known as "data management problems". You create your own infrastructure to run the analysis as far as it needs a home-made mixing of tools, then you extract the data, start running the scripts, and so forth.
Soon enough, you have created a mess of random of droppings, and if you come back to this session two weeks later, you don't recall what the files meant, and so forth.

5. A Way for Mixed Strategies

In accordance with the type of problems presented before and just as an example, some running systems from these industries are presented here.
Firstly, we present a system to "estimate" the mechanical properties of galvanized coils. The main goal was to optimize the quality control of galvanized steel by developing a predictive model of the mechanical properties according to the chemical composition and manufacturing conditions in the annealing furnace [10].

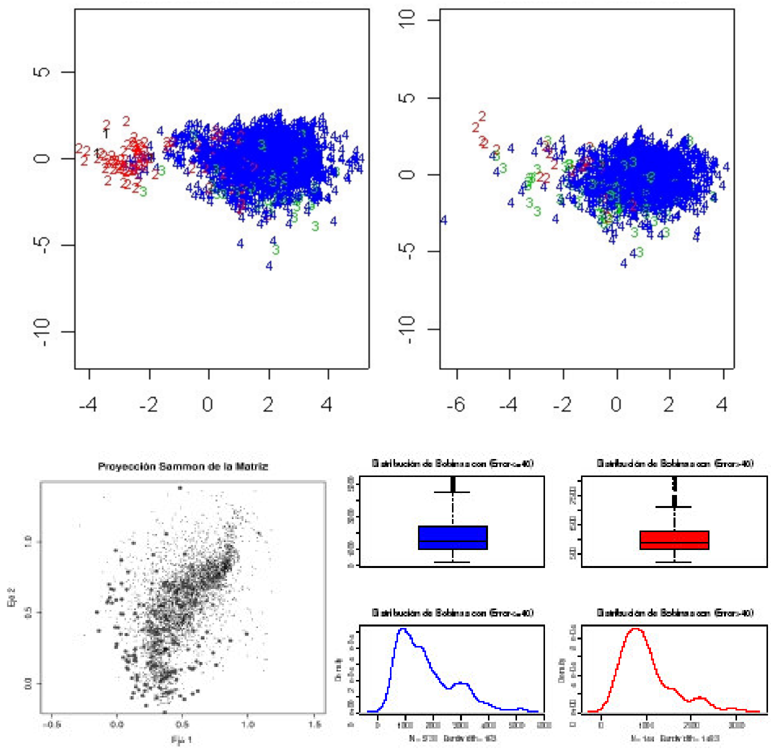

Figure 7. Using Multidimensional Projectors in order to identify outliers.

In order to produce high-quality galvanized steel economically, it is necessary to control the processes' conditions to fulfil all quality demands. There have been improvements in the modelling and control of the annealing furnace, in the coating system control, etc.

In order to form a continuous strip, coils are uncoiled and a shear cuts off the end of each coil so that they can be welded together. Then, the oil, dirt and oxides on the surface of the cold-rolled coils are removed before the strip enters the annealing section of the line. A good adherence, necessary to obtain an excellent coating quality, is achieved by a perfect strip cleaning.

The clean strip passes through the annealing furnace to give the steel the desired properties by heating it to a particular temperature profile that determines the grain structure within the metal and prepares it for the galvanizing process. The entire process is carried out in a protective atmosphere that also reduces the surface of the strip used in the coating preparation step. The annealing cycle includes the following phases:

1. The cold strip is recrystallized by bringing it to the highest temperature of the annealing profile.
2. The strip temperature is maintained and the expansion of the grain takes place.
3. A slow cooling is used to control the metal texture.
4. Then, a fast cooling prepares the steel for the strain-aging treatment. The strip is cooled to a temperature appropriate for the coating stage.
5. The overaging step results in the precipitation of carbon such that the carbon solute is reduced. Thus, the strain-aging tendency of the strip is reduced.

After the annealing step, the strip enters the molten zinc bath in order to form a zinc coating that is metallurgically bonded to the steel surface. The coating thickness is controlled by air currents positioned after the zinc bath. The control of the

coating thickness is one of the most critical areas of development for coated sheets.
Finally, the coated strip is subjected to a chromate conversion treatment by the application of chromate solutions to the strip surface. This chromate treatment results in a surface resistant to corrosion during storage and transport until steel can be used in other applications.
Presently, the mechanical properties of galvanized sheets and coils are measured after their fabrication. Due to the off-line control, a large dead time passes, which makes the control solution inefficient. That is, the continuous galvanizing line produces at least two coils or sheets from the very moment a coil with undesired properties is detected until appropriate actions are taken. Such a delay results in the cost of a coil, whose quality is sub-standard.
In order to improve the control system and allow for the on-line control of the desired mechanical properties, we used data mining and artificial intelligence techniques to develop a predictive model. With this model, the impact on the manufacturing conditions in the annealing furnace on the final mechanical properties was analyzed.
The aim of the data mining analysis was to predict the yield strength, tensile strength and elongation as functions of a large number of variables, including the chemical composition, heat treatment and strip speed in the annealing furnace. Missing values were not allowed since they made the predictions noisier.

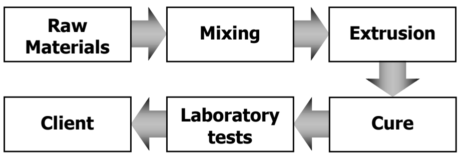

Figure 8. Rubber process.

Another case for further illustrating the previously explained concept relates to the rubber industry, since there is no certain knowledge of the relationship between the production conditions and the final properties of the extruded product even if the production process involved has been used for a long time. Very serious efforts are being made both in rubber mixing and extrusion processes. A number of works in the specialized literature can be found [1].

Some works aimed at creating models to predict Mooney viscosity, cure characteristics, etc.; have already been published. This work extends the model, where the minimum and maximum torque and the time to reach 50% of the cure-state were modelled using compound formulations. Our approach also takes into account mixing conditions to predict other important parameters of the rheometric curve, such as the scorch time and cure time.

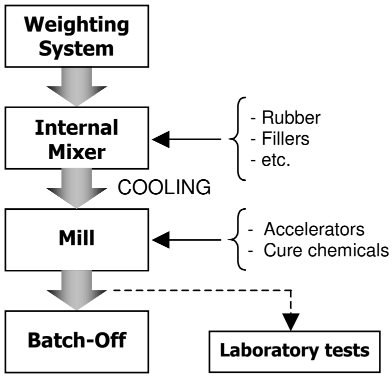

Figure 9. Adding ingredients into rubber process.

By integrating the proposed models into the manufacturing system, it is possible to improve the information flow through the plant as it is no longer necessary to wait for laboratory results. Thus, product traceability is guaranteed and scheduling and control functions are improved as real time information enables us to make control decisions that hold to the production plan. Some authors proposed a generic framework to facilitate the design of any specific monitoring environment and to allow for the integration of optimization tools in the manufacturing system.
These improvements mean that the quality of the extruded product can be increased and the scrap rate and failure costs can be decreased as they allow for the on-line planning of manufacturing orders according to the specific characteristics of the blends.
Rubber mixing is performed in batches and involves two phases in order to avoid a premature cure of the blends due to the high temperatures reached during the mixing process.

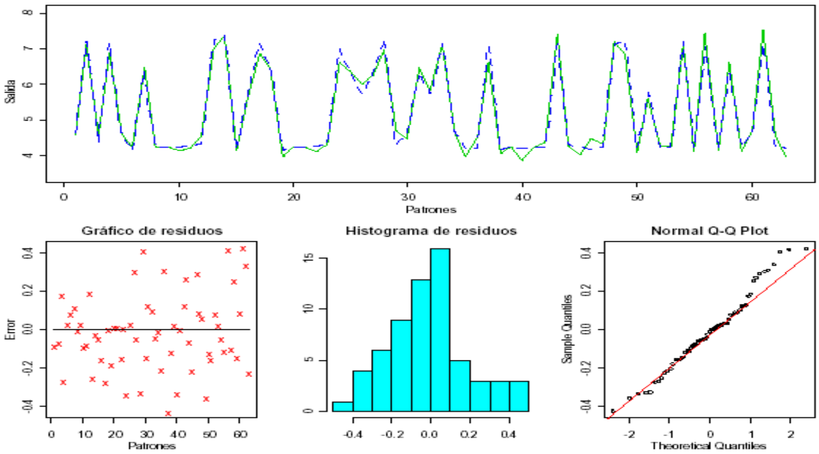

Figure 10. Neural Network Training Results.

First, the polymer, together with carbon black and other fillers, plasticizers and additives, is fed into an internal mixer. Ingredients are added during two or three stages, depending on their quantity. Upon completion of the mix cycle, the mixture is discharged through a door at the bottom of the mixer.
Once the batch is cooled, it is dumped onto a mill where it is further worked while being cooled. Additional compounding ingredients, such as curatives, are added at this point. Finally, a mixed rubber sample is taken to perform a test and the blend passes through the batch-off machine where it is cooled, cut and stored.

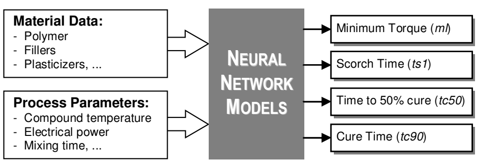

Figure 11. Neural Network Applied.

The rotor speed and mixing time were kept out of the analysis as they were the same for every blend in each compound and so they were useless. Therefore, each model consisted of 11 input variables and 4 output variables. A linear regression analysis was also used to obtain the model of the cure characteristics. Again the results were not good. Then, a neural network analysis was applied to estimate the minimum torque, scorch time, time to 50% cure and cure-time of the blends.
To summarize, in this work [11], we have developed four different models: for each cure characteristic a linear model, a neural network with 11 inputs, a neural network with 7 inputs, and a neural network whose inputs were the projection obtained by PCA.
It has been shown that it is possible to create reasonable neural network models for the cure characteristics analysed, considering the compound formulation and mixing conditions. In this work, we have shown the results obtained for one compound. However, this methodology can be applied to other rubber formulations mixed with any rotor type.
This technology can be extended to the industrial maintenance field as far as we can identify 'tendencies' in several parameters and to identify main reasons for these parameter deviations. This strategy allows to people managing the maintenance policies, to decide the best procedural preventive maintenance, taking into account not only the process itself but the quality of the final product

6. Conclusions

The data mining technology shows excellent capabilities to be used for several purposes, including:

- Product quality monitoring, specially for those characteristics related to continuous product production.
- Process monitoring as a key aspect for previous goal.
- Maintenance strategies as a very relevant aspect related to the product quality and to the economical performance of the plant

Also it seems to be clear that Data Mining can be improved by using numerical modelling simulation where not real data are available or where an initial parameter calibration is required in order to operate the system.

In all cases, the impact of these technologies in modern industrial processes makes a requirement for companies to be aware of them in order to extract as knowledge as possible from their processes and be able to improve them.

Acknowledges

We will grateful the INTERREG III-A programme for supporting partially this initiative. Also we express our thanks to the DPI2004-0762-C2-01 project funded by the Spanish Education Ministery, to the RFS-CR-04023 EU funded research programe and also to the II Plan Riojano de I+D for its continuous support and help.

References

[1] González A., Pernía, A., Alba, A. García. A neural network based approach to optimize rubber extrusion lines. Sent for publishing to Int. J. Computer Integrated Manufacturing.

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

[3] Bill Kahn, 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.

[4] Castejón, Ordieres, Martínez de Pisón & Vergara. Outlier Detection and Data Cleaning in Multivariate Non-Normal Samples: The PAELLA Algorithm. Data Mining and Knowledge Discovery, 9, 171-186, 2004.

[5] David Duling, SAS. Computational Performance in Data Mining. M2004, the 7th annual Data Mining Technology Conference. Las Vegas. USA.

[6] Espinoza, J. Ordieres, F.J. M. de Pisón, A. G. Marcos. Tao-Robust Backpropagation Learning Algorithm. Sent for publishing to "Neural Networks " journal.

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

[8] Hawkins, D. Identifications of Outliers. New York: Chapman and Hall. 1980.

[9] Jim Georges, SAS. Using non-numeric data in parametric prediction. M2004, the 7th annual Data Mining Technology Conference. Las Vegas. USA, 2004.

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

[11] Ordieres, J.B., López, L.M., Bello. Intelligent methods helping the design of a manufacturing system for die extrusion rubbers. International Journal of Computer Integrated Manufacturing. 16 : 173-180, 2003

[12] Michael A., Boeing. Cooperative data mining: Tightly integrating data mining with visualization. M2004, the 7th annual Data Mining Technology Conf. USA, 2004.

[13] 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).

[14] Rocke, D. and D. Woodruff, Identification of outliers in multivariate data. J. Amer. Statist. Assoc. 91, 1996. pp 1047–1061.

[15] Rocke, D. and D. Woodruff. Robust Estimation of Multivariate Location and Shape. Journal of Statistical Planning and Inference 57, 1997. pp 245–255.

[16] Rousseeuw, P. J. and A. Leroy, Robust Regression and Outlier Detection Diagnostic Regression Analysis. New York: John Wiley and Sons. 1997.

[17] Srivastava, M. S. and D. von Rosen, Outliers in Multivariate Regression Models. Journal of Multivariate Analysis 65, 1998. pp 195–208.