

# ANN Quality Diagnostic Models for Packaging Manufacturing: An Industrial Data Mining Case Study

Nicolás de Abajo  
Aceralia Arcelor Group  
Innovation and Research  
Nicolas.Abajo@arcelor.com

Vanesa Lobato  
Aceralia Arcelor Group  
Innovation and Research  
Vanesa.Lobato@arcelor.com

Alberto B. Diez  
Department of Electrical Engineering  
University of Oviedo

Alberto@isa.uniovi.es

Sergio R. Cuesta  
Department of Electrical Engineering  
University of Oviedo

srcuesta@isa.uniovi.es

## ABSTRACT

World steel trade becomes more competitive every day and new high international quality standards and productivity levels can only be achieved by applying the latest computational technologies. Data driven analysis of complex processes is necessary in many industrial applications where analytical modeling is not possible. This paper presents the deployment of KDD technology in one real industrial problem: the development of new tinplate quality diagnostic models.

The electrodeposition of tin on steel strips is the most critical stage of a complex process that involves a great amount of variables and operating conditions. Its optimization is not only a great commercial and economic challenge but also a compulsion due to the social impact of the tinplate product—more than 90% of the production is used for food packaging. The necessary certification with standards, like ISO 9000, requires the use of diagnostic models to minimize the costs and the environmental impact. This aim has been achieved following the multi-stage DM methodology CRISP-DM and a novel application of pro-active maintenance methods, as FMEA, for the identification of the specific process anomalies. Three DM tools have been used for the development of the models. The final results include two ANN tinplate quality diagnostic models, that provide the estimated quality of the final product just seconds after its production and only based on the process data. The results have much better performance than the classical Faraday's models widely used for the estimation.

### Categories and Subject Descriptors:

I.5.2 [Pattern Recognition]: Design Methodology

**General Terms:** Algorithms, Management, Design.

**Keywords:** Tinplate quality, ANNs, CRISP-DM, FMEA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

KDD'04, August 22–25, 2004, Seattle, Washington, USA.

Copyright 2004 ACM 1-58113-888-1/04/0008 ...\$5.00.

## 1. INTRODUCTION

Steel is one of the leading materials used for attractive, convenient and cost-effective packaging today. There are several types of packaging steel for different container applications and the most commonly used is the tinplate. Its name derives from the thin layer of tin that is electrolytically applied to a low carbon steel base. In this process the steel strip passes through several tanks increasing its tin coating thickness. These tanks contain the electrolyte and the pure tin anodes so when there is a voltage difference between the tin anodes and the steel cathode, the tin plating takes place.

The thickness of this microthin layer of tin is crucial. It depends on the type of application varying from the lowest values  $-0.9 \text{ g/m}^2$  to the highest ones  $-14 \text{ g/m}^2$ . Too low tin coating thickness accelerates corrosion which may spoil the food packed in steel cans with undesirable consequences for public health. On the other hand, the effects of too high coating thickness are not so dramatic but it makes difficult the following stages of production and increases the cost dramatically.

However, coating thickness measurement for diagnosis presents many difficulties. The on-line systems, based on radiographic techniques, provide a precise measure of several features along the strip but they have two main problems: they are very expensive and they require many safety and maintenance checks.

So being the coating thickness a critical parameter of the tinplate quality requirements, this paper propose a novel quality diagnostic model that meets the new ISO-9000 standards and it is just based on the process data with no other additional measurements.

The paper is organized as follows. In section 2, a brief description of a tinplate line is given. Then, it is introduced the necessity of control of the coating thickness, the currently on-line measurement systems and the different existing models. Section 3 shows in detail the data mining process carried out for obtaining the new models and following the multi-stage methodology CRISP-DM<sup>1</sup>. Finally, section 4 concludes the paper.

<sup>1</sup>Cross-Industry Standard Process for Data Mining

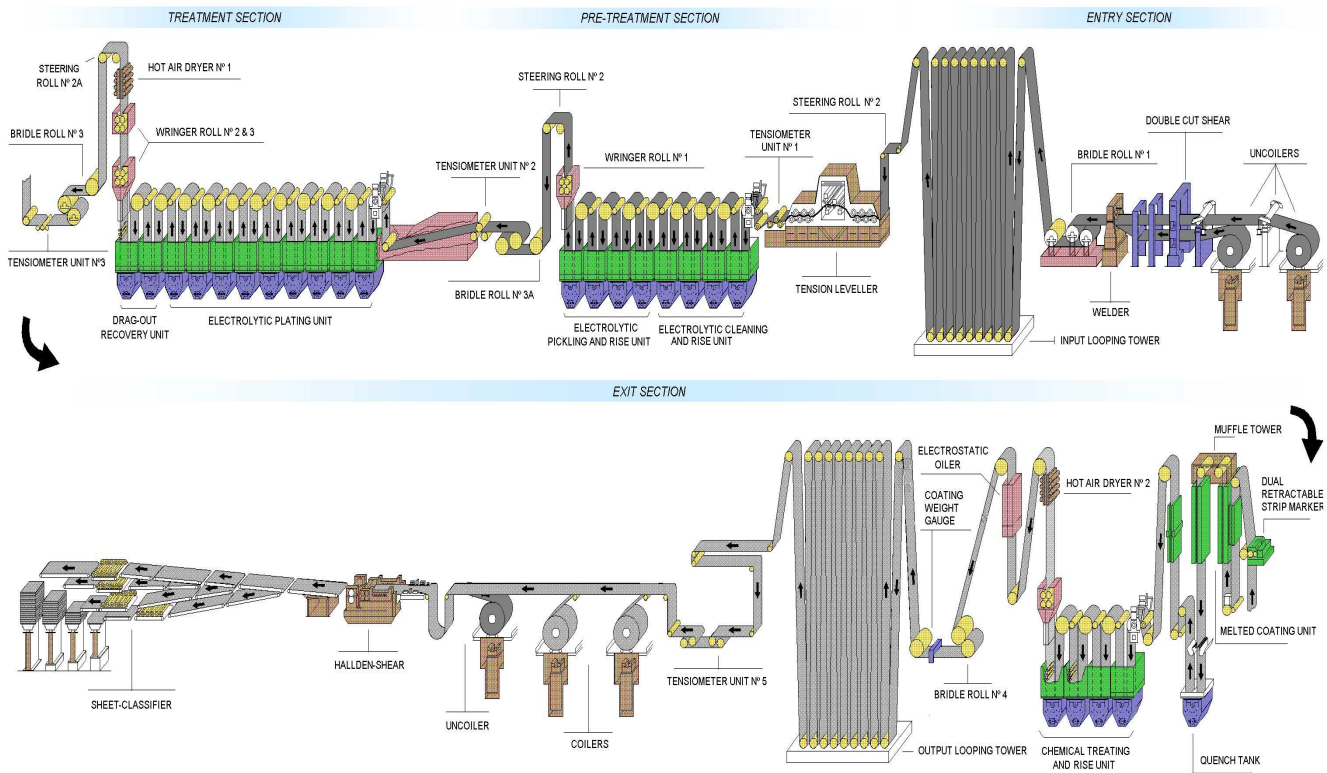


Figure 1: Tinplate Line Layout.

## 2. TINPLATE PRODUCTION PROCESS

Currently, tinplate is used for a variety of consumer and commercial applications. It offers particular advantages for packaging such as strength, workability, corrosion resistance and weldability. It is also environmental friendly being completely recyclable and having an energy-efficient production process [1].

### 2.1 Description of a tinplate line

1. **Entry section** which levels the coil through a combined action of draw and bend, and makes the process continuous, by welding the head of a coil with the tail of the previous one, see figure 1.
2. **Pre-treatment section** which realizes the cleaning of the coil through an electrolytic alkaline degreasing and an electrolytic acid pickling, necessary because the coil is usually dirty with oil, grease and solid particles coming from the rolling process.
3. **Treatment section (tinplating)** with nine vertical electroplating cells which increase the thickness on the coil surface at each step, thanks to an electrolytic process, until the desired layer thickness is obtained when exiting the last cell.
4. **Exit section** with a differential marking treatment, useful to determine the side with higher tin coating thickness, a melting-brightening treatment which gives the tin coating a bright surface, better resistance and forms the Fe-Sn alloy that gives the sound adhesion to the base metal and a electrostatic oiling which provides

lubrication to the coil surface to minimize abrasion damages and to facilitate mechanical operations [2].

### 2.2 Coating Thickness Control

In general, metallic coating is defined by its thickness, porosity, adhesion and resistance to service conditions. In the case of tinplate, the tin coating determines both the capacity of isolation and the following treatments of the material (lacquers, oils, etc.). These treatments improve the external image of the product and protect the contents from the environment.

In the past decades, average tinplate thickness has been reduced both for economic and environmental reasons increasing the process complexity. An important effort in terms of R&D devoted to the tinplate production is focused on improving the coating thickness applied.

At the same time, industrial enterprises are increasingly aware of the need to improve product quality and productivity developing and implementing appropriate quality management system solutions. New edition of ISO 9001:2000 [3], increases emphasis on the determination and use of not only statistical methods as SPC (Statistical Process Control) but also predictive based on:

- Planning of product realization: including the required verification, validation, and inspection and testing processes specific to a product.
- Planning of the monitoring, measurement, analysis and improvement of process and products to demonstrate the ability to meet requirements.

Radiometric gauging is the most widespread option for on-line measurement of coating thickness. This kind of gauges operate by means of gamma radiation. Their main drawbacks are the impossibility of locating the obtained measurements with a high degree of accuracy, due to the off-center strip movements, their high price and its requirement of frequently calibration.

Most of the approaches to improve the final coating quality are based on the Faraday's law of electrolysis. It states that the amount of material deposited in the cathode (steel strip for us) in the electrolysis process is proportional to the electrical charge and the equivalent weight of the substance (tin in our case). Due to other non-metallic substances (like Hydrogen) the final amount of substance on the strip is not the theoretically expected.

The thickness model takes into account all these factors in one  $K$  with a set value for each thickness. The given formula is  $I = K \times a \times v$ , in which  $I$  is the current in the rectifiers,  $v$  is the line speed,  $a$  the strip width and  $K$  the coating factor.

Other models try to overcome the excessive simplicity of the Faraday's law with the integration of the main principles of the mass movements equations. These methods manage extremely complex equations, usually solved with FEM, so are not an adequate answer to the plan necessity for quick diagnosis.

### 3. DATA MINING PROCESS

This section describes in detail each phase of the Data Mining process that was carried out following the multi-stage methodology CRISP-DM [4].

#### 3.1 Business understanding

**Previous knowledge.** The most important factor for the coating is the current applied. At present, it is determined based on the customer order data, by applying the generic Faraday formula with a specific factor fixed experimentally. The advantages of this system are its simplicity, calculation speed and excellent expert knowledge. The disadvantages are that the system is not adaptive and it does not take into account all the influences that are relevant to the problem.

**Business objectives.** Classification of coils by average coating thickness, based on process conditions. This provides an increased quality for customers with very specific requirements, increased productivity thanks to higher process speeds, improved environmental performance, reduced costs thanks to a reduction in the amount of tin deposited on the strip and decrease rejections.

**Success criteria.** 95% of the coils classified within the correct range. The determination of the relevance of certain process variables which have a notable environmental impact (cleaning and pickling) in order to minimise them as well as better capability to meet certification requirements.

**Resources.** On-line process data recorded at three different process levels. The sources of knowledge used are the expert knowledge of the technical team and the maintenance staff, the performance reports, the description of the production line and the FMEA<sup>2</sup> analysis of the facility.

**Risks and contingencies.** There are both market-related risks, resulting from competition, and technical risks, mainly

due to the potential unreliability of some of the data and to the many factors external to the line that influence the process, as was observed in the FMEA analysis [5], [6].

Another key issue is the communication between the various operational levels, where data synchronization is an essential requirement.

**Benefits.** The implementation of a new *formula* for calculating the setpoint enabling the excess coating to be reduced by 30% could lead to savings estimated at 0.5 MEuros/year. The indirect benefits include minimisation of the environmental impact and an enhanced understanding of the production process.

**DM objectives and success criteria.** For 95% of the cases, an error of less than 10% in the predicted average quantity of tin deposited on the coil, for given conditions of relevant variables. The success of the project will be assessed by comparing two models: the classical Faraday model and the model adjusted by the technical team.

**Planning.** The task distribution entailed about two years' work. The critical points were the definition of objectives, proper data understanding and collection, adequate preparation and classification of the operating standards for training the models, and a realistic implementation on-site.

**Tools and techniques.** From among the various DM tools available, the solution selected was *Data Tools*, based on versatility and cost criteria. It is a tool developed by BFI<sup>3</sup> supported on *Matlab* toolboxes and proprietary algorithms.

#### 3.2 Data understanding and preparation

**Data acquisition:** There is a data log corresponding to six months' production, amounting to 3Gb of information. The selection criterias were:

- Process: to avoid irrelevant or post-coating variables.
- Delphi methodology: interviews with the members of the technical teams to assess the relevance of the variables with regard to the objectives defined.
- The FMEA document, in order to avoid including in the analysis alterations caused by other facilities.
- Mathematical techniques for variable selection and plausibility checks.

**Data description:** From all the variables available, the proposed list of 21 most relevant ones, drawn-up in collaboration with the facility staff, is the following:

- Heat-exchangers and electrolyte re-circulation temperatures and levels in tanks 1 and 2: TEMP\_E1, TEMP\_T1, LEVEL\_T1, TEMP\_E2, TEMP\_T2, LEVEL\_T2.
- Voltage and current in each block of rectifiers: CURR\_T, CURR\_B, VOLT\_T, VOLT\_B.
- Line speed and strip width: SPEED and WIDTH.
- Actual and target strip coating thickness (Top and Bottom side): COAT\_T, COAT\_B, TARG\_T, TARG\_B.
- Electrolyte concentrations: conductivity, COND, and pH.
- Upstream data from cleaning and pickling sections: CURR\_C, VOLT\_C, CURR\_P.

<sup>2</sup>Failure Mode Effect Analysis

<sup>3</sup>Betriebsforschungsinstitut GmbH

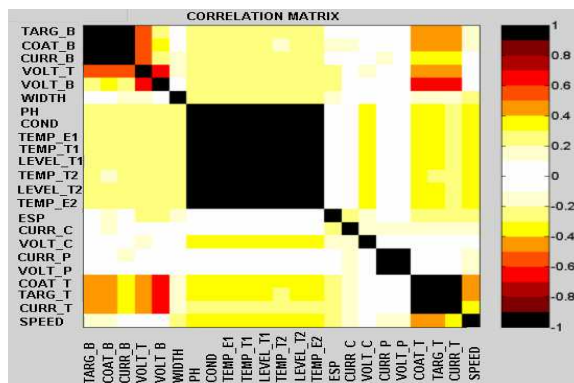


Figure 2: Correlation Matrix.

**Data quality verification:** In order to study the plausibility of the data it is necessary to apply a series of physical ranges for the key variables. The main alarms are those related to concentrations, temperatures and currents.

A particularly important point is the evolution of the current variable as a function of its setpoint values [7]. It shows a certain hysteresis that prevents a linearization between the current setpoint value and the measured one. The variation observed in the data involved in this relationship (which should be linear) led to the implementation of a contingency plan in collaboration with the maintenance department, resulting in the launch of a new data acquisition campaign.

After a preliminary analysis of the data, the TEMP\_E1 variable was rejected in view of the low quality of the input data. The physical plausibility tests for the rest of the variables were valid.

**Data exploration:** The basic statistical analysis of the data shows the suitability of the study, in view of the differences existing between the target and the actual thickness measured, the low quality of the data for the voltage-related variables and the limited variability observed in some of the electrolyte variables [8], [9].

Figure 2 shows the high degree of correlation existing between some of the data that characterise the electrolyte: these are basically the electrolyte concentrations, temperatures and levels. A high degree of correlation is also observed between the actual and target coating, and between the current applied on each side of the strip. Logically, there is also an inverse correlation between the coatings applied on the top and bottom side of the strip and the speed of the line.

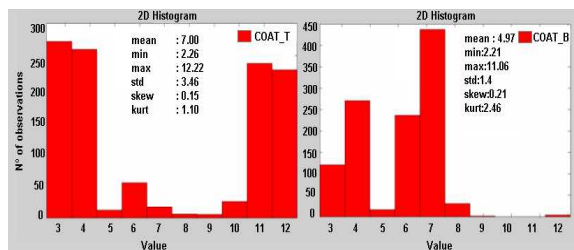


Figure 3: Target variable histogram.

The identification of data sets is based on the analysis of the target variables TARG\_T and TARG\_B. As can be seen in figure 3, the distribution of coating thicknesses varies with the

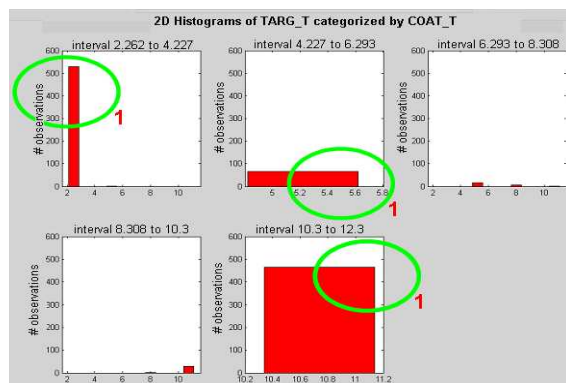


Figure 4: Target vs actual coating histogram.

strip side: medium and low coating thicknesses are mostly observed on the bottom side whereas low and high ones are more predominant on the top side. This implies that the major operating limitations of the line (speed) are usually determined by the setpoint values of the top side.

In figure 4, the histograms of the target variable TARG\_T are displayed according to the equivalent histograms of the variable COAT\_T classified according to five groups generated with *C-means*. The circles indicate the three zones of low, medium and high coating thickness observed in the other variable.

This type of analysis applied to different groups of variables led us to the following conclusions:

- Coherence between actual and target coating and the current values.
- pH, conductivity, temperature, evaporators and the variables from tank 1 show an analogous distribution, with lower values for thicker coatings.
- The voltages related to COAT\_T are, on average, higher than those related to COAT\_B.
- High values of COAT\_T entail low values of COAT\_B, low speeds, wider formats, inferior cleanliness and thicker coatings. Thinner coatings entail higher cleaning currents.

The clustering tasks required for downstream tools, *GAs* and *DTs* algorithms applied deal with discrete target variables, were also carried out in this stage. By using a knowledge processing similar to that applied in the Delphi methodology, a classification into three groups of high, medium and low thicknesses homogeneous in terms of volume and with regard to the three basic variables (current, target values and speed) was achieved.

**Mathematical selection of relevant variables:** The *Self-Organizing Map (SOM)* method is a powerful algorithm for the visualization of high-dimensional data [10], [11]. Particularly insightful are the so called *component planes* represented in figure 5, that provide us with a *big picture* of the input values distribution. Similar maps show an analogous behavior and, therefore, a redundancy in the information. This leads to the following conclusions: TARG\_T and COAT\_T are projected in the same way in the visualization space. The same applies to COND, TEMP\_E1, TEMP\_T1, LEVEL\_T1 and TEMP\_E2 and to pH, TEMP\_T2 and LEVEL\_T2.

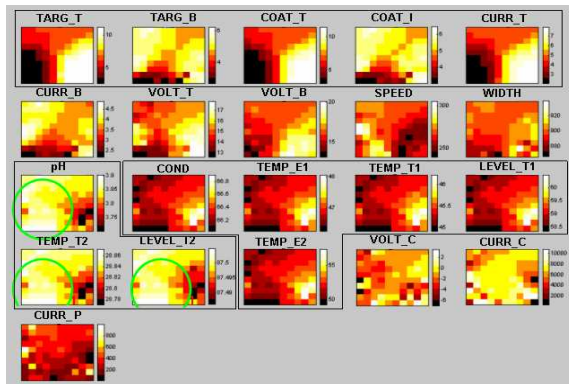


Figure 5: SOM component planes.

For the variable selection criteria, the three main sub-groups created were used, and genetic algorithms (*GAs*) and decision trees (*C4.5* & *OC1*) were applied. The aim is to identify the combination of variables that contain more information about our target variables and subsequently those that show the best classification performance. The algorithms results are shown in figure 6 as a histogram with the most frequently selected variables for growing the trees and the most important ones used in the *GAs* individual input combination, thus leading to the following 12 key variables: CURR\_T, CURR\_B, TARG\_T, TARG\_B, SPEED, WIDTH, pH, CURR\_C, VOLT\_T, VOLT\_B, TEMP\_E2, CURR\_P.

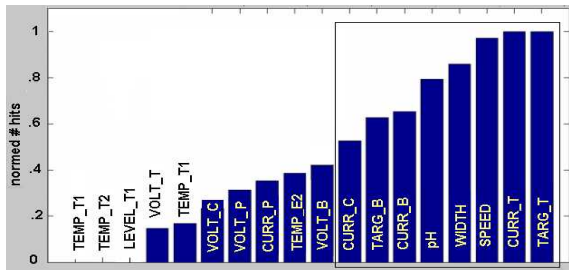


Figure 6: *DTs* & *GAs* algorithms variables selection.

Starting with several dozens of variables, the problem was reduced to twelve inputs. Having selected the relevant variables, the available cases are filtered according to their relevance for the analysis eliminating tuples that affected by some contained in the FMEA document. With the above restrictions, the number of cases to be processed is reduced to a data set of about 1300 coils which means a 50% reduction on the initial number of cases.

### 3.3 Coating Modeling

The desired output of the model is a forecast of the tin coating, enabling us to determine, for any condition that may arise in the production line, whether the final coating will be excessive, adequate or insufficient.

**Description:** The plan for the design of new models included the following phases: in a first phase, we worked with all the variables requested by the technical team (21); then, we eliminated the variables showing a linear correlation greater than 0.99 (18 variables remained); next, the variables previously considered as not highly relevant were

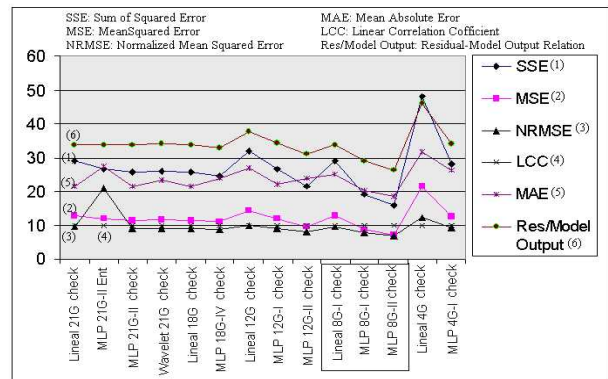


Figure 7: Results for different models.

eliminated, leaving 12 and then the first 8 variables of figure 6. Finally, a minimal model is drawn-up with the 4 inputs of Faraday’s law.

Using this sequence of (21/18/12/8/4) inputs, we start with a linear regression model that serves as a base comparison, evolving to Multilayer Perceptron (MLP) models with Bayesian learning features and sigmoidal activation functions. The whole data set is balanced split: 50% for training, 25% each for validation and testing. Then, the models obtained are run on 250 coils that cover the whole product-mix, enabling us to reproduce their behaviour under real conditions. Figure 7 shows the evolution of the errors rates for some cases. Those models calculated using the eight inputs preselected in section 3.2 show the best performance in terms of error rates. Then, we look through them.

**Models:** Figure 8 shows the results obtained from checking the linear model. As it can be seen in (1), the average value is not bad, but shows very high residues (2) and a single output for high values (3). Then, MPL models were formulated. The results achieved when checking the MLP 8G-I model (see figure 9) are very promising and show a 35% improvement on the linear model; in this case, the problems lie in the presence of residues of very high values (2) and, especially, in the area of low coating thicknesses outside the safety band (3) [12].

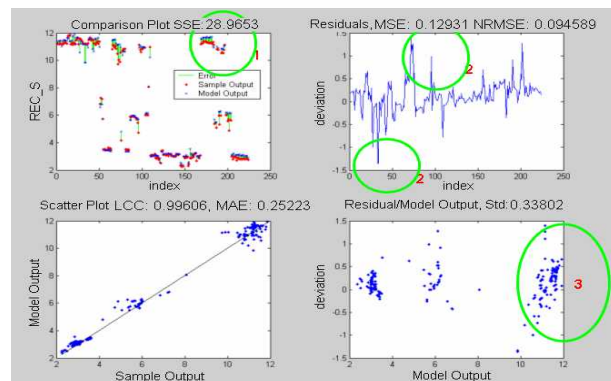


Figure 8: Linear 8G-I model check results.

In view of the good results achieved with the previous model, tests were performed using another learning algorithm. The result in average values (see figure 10) was good:

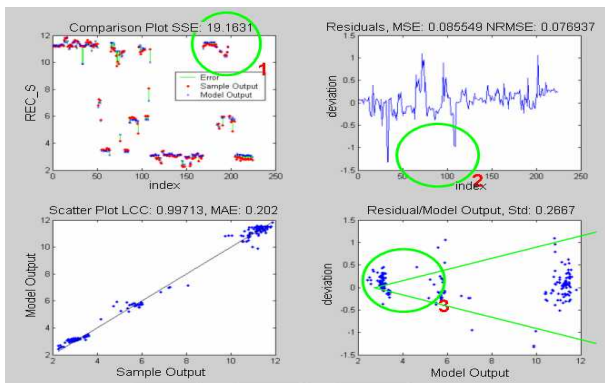


Figure 9: MLP8G-I model check results.

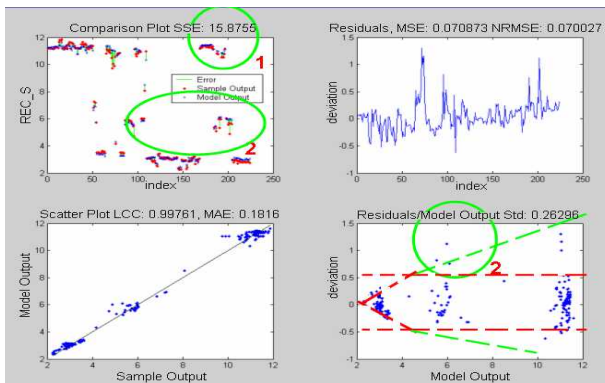


Figure 10: MLP8G-II model check results.

an 85% improvement on the performance of the linear models and a 21% improvement on the best previous model. It explains 99.7% of the cases (LCC) and its behaviour in terms of absolute and relative residues is optimal, enabling us to configure the forecast in a much narrower range than originally targeted for medium and high coating thicknesses. The only drawback appears in three coils with medium coating thickness, where problems arise. In view of the high performances of the model, we study these specific cases.

Among the average coating thickness verification variables, a series of atypical cases were observed in the current variable and these cases coincide with the maximum errors of the model. These three values, which lie outside the system logic were the cause of the outliers of the model. The results obtained for the bottom side coating model are very similar to that of the top side coating, as it could be expected [12].

Finally, the model proposed by *Data Tools* was validated by comparing it to two other tools, namely *Clementine* and *MARS*, for the selection of variables and modeling. The model proposed is also a MLP with the same selection of variables. A high degree of consistency was shown in all the results obtained.

#### 4. CONCLUSIONS

The main conclusion to be drawn from this work is the development of two tinsplate quality diagnostic models (one for each side) that provide the estimated quality of the final product just based on the process data. Both of them are

multilayer-layer perceptron networks with Bayesian learning and sigmoidal activation functions. These models improve the accuracy of classical Faraday's linear models explaining 99.7% of the cases and showing a good performance in terms of absolute and relative residuals. The economical impact of this model is, beyond any doubt, very important.

Another excellent output has been to show the relevance of a set of variables, not taken into account before, by means of the application of both classical and advanced statistical data analysis techniques and other structured processes for collecting and distilling knowledge. The list of relevant variables was also a combination of expert knowledge provided by maintenance (FMEA) and production experts. Finally, this aim has been achieved following the CRISP-DM methodology and according with new ISO 9000 standards.

#### 5. REFERENCES

- [1] E. de la Toba. *El Proceso Siderurgico*. ACERALIA, 1998.
- [2] V. Ferrari, F. Sanfilippo, N. Di Biase, E. Musella, N. de Abajo, J. A. González. System for Electroplating process diagnosis based on virtual sensors and internet technology. In *7th Tinplate Conference*, Amsterdam, 2-4 October 2000.
- [3] Iso 9000 international standards for quality management. *International Organization for Standards*, 1991.
- [4] P. Chapman, J. Clinton, J. Hejlesen, R. Kerber, T. Khabaza, T. Reinartz, and R. Wirth. The current crisp-dm process model for data mining, 1998.
- [5] D. Stamatis. Failure mode and effect analysis. 1996.
- [6] S. Keplinger. A new plant-wide system for quality control. In *Milenium Steel 2k2*, pages 290–294, 2002.
- [7] N. de Abajo, S. Peregrina, J. A. González, and A. López. Development and implementation of intelligent systems for the real time optimization of product quality in rolling mills and process lines. In *ECSC Workshop: Applications of Artificial Intelligence in Real Time Applications*, pages 181–195, Brussels, 2000.
- [8] H. Peters, T. Heckenthaler, and N. Holzknicht. Optimisation of flat product quality by intelligent data exploitation. In *Proceedings of 3rd European Rolling Conference*, pages 317–322, Dusseldorf, 2003.
- [9] H. Peters, T. Heckenthaler, and N. Link. Application of data mining methods to find correlations between quality data and process variables. In *Proceedings of 10th IFAC Symposium on Automation in MMM*, pages 141–146, Tokyo, Japan, September 2001.
- [10] S. R. Cuesta, I. Díaz, A. A. Cuadrado and A. B. Diez. A visual approach for fuzzy rule induction. In *9th IEEE International Conference on Emerging Technologies and Factory Automation*, pages 761–767, Lisbon, Portugal, 2003.
- [11] A. A. Cuadrado, I. Díaz, A. B. Diez, F. Obeso, and J. A. González. Visual data mining and monitoring in steel processes. In *37th Annual IEEE Industry Applications Society (IAS) Meeting*, pages 493–500, Pittsburgh, PA, USA, 2002.
- [12] N. de Abajo. *Optimización mediante Data Mining de modelos para el diagnóstico de calidad en hojalata*. PhD thesis, Universidad de Oviedo, 2001.