A Data Mining Analysis Applied to a Straightening Process Database[1]

Jean-David Caprace, Nicolas Losseau, Frederic Bair, Philippe Rigo, University of Liege[2]
Dominique Archambeau, PEPITe S.A.

1 Introduction

The complexity of modern manufacturing processes in a highly competitive environment forces the manufacturers to invest massively in automation and monitoring systems. The large data flows from these new installations are sources of valuable and hidden knowledge that is so far hardly used. Data mining methods through integrated data analysis tools give a solution to this situation, allowing easy retrieval of knowledge starting from a data base. This is also a unique opportunity to learn faster about the process and to detect hidden and complex relationships between parameters involved. Within this framework we have decided to apply this data analysis method to the straightening process in shipbuilding. We refer to *Caprace et al. (2007)* for additional illustrations.

In shipbuilding, the assembly of elements by welding involves temperature gradients within the material. These cause deformations which sometimes have to be reduced to obtain an acceptable surface flatness. The straightening process to eliminate these distortions for esthetical or functional reasons is labour intensive. Estimating the straightening impact on the production workload is interesting in the context of production simulation, cost assessment of ship hull, structure optimization, design for production, etc.

To reach these objectives, the idea was to elaborate, through a data mining approach, a formula linking the straightening cost to the sections scantlings (plate thickness, dimension and inter-distance of longitudinal stiffeners, dimension and inter-distance of transversal frames) and to other section characteristics. This paper describes each stage of the methodology: data description, analysis of data quality, data exploration and finally choice of discriminatory attributes and the generation of the data-driven models.

2 Data Mining

Data Mining (DM) is aims to extract synthesized and previously unknown information from large databases. By definition, DM extracts valid, previously unknown, comprehensible, and useful information from large databases for further processing. The data analysis process tries to discover useful patterns in data that are not obvious to the user and to create useful information and knowledge from data to help in decision-making.

DM is increasingly used in enterprises decision-making processes. DM uses a broad range of tools from statistics, automatic learning, pattern recognition, database technologies, visualization and artificial intelligence. This mix of technologies was combined in the PEPITo software, `www.pepite.be`, used in the work presented here. PEPITo combines visualization, advanced statistics and predictive analytics tools. Its scalable, high-performance object database allows to handle problems with very large data sets (millions of objects with thousands of parameters each). The simple and intuitive user interface hides the complexity of DM from users, enabling them to quickly perform DM tasks on simple problems. In addition, PEPITo has a powerful scripting language (KDLisp). KDLisp is fully programmable with direct access to all the DM functions, allowing a flexibility which is important, since most DM applications involve an iterative process of exploring data models with different data mining techniques. The toolbox concept enables users to easily prototype different data models and

[1]Research funded in part by the sub-project II.1 of the IP INTERSHIP Project, the MARSTRUCT project and the STREP IMPROVE Project (031382-FP6 2005 Transport-4).
[2]ANAST, Chemin des Chevreuils 1, BE-4000 LIEGE 1, Belgium, jd.caprace@ulg.ac.be

to combine, modify and adapt the tools according to characteristics of the problem. It also allows users to integrate the solution with other software.

3 The data mining process

We followed the freely available CRISP-DM data mining methodology, `www.crisp-dm.org`. This methodology was developed in the 1990s by a consortium of DM experts, consultants and end users. The proposed reference model covers all aspects of a DM study, from problem definition, to data analysis and final live deployment. We briefly summarize the different stages of the process below:

1. Business Understanding – This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.

2. Data understanding – The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

3. Data preparation – The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modelling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include tables, objects and attributes selection as well as transformation and cleaning of data for modelling tools.

4. Modelling – In this phase, various modelling techniques are applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same DM problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to data preparation phase is often necessary.

5. Evaluation – At this stage in the project, you have built a model that appears to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to thoroughly evaluate the model and review the steps executed to construct the model to be certain that it properly achieves the business objectives.

6. Deployment – Depending on requirements, the deployment phase can be as simple as generating a report, or as complex as implementing a repeatable DM process across the whole enterprise.

4 Straightening process database study

Various tools for data analysis were used in this study in order to find a relevant relation between the straightening cost and the sections scantlings. As a first step, a database was established in order to specify the characteristics of each section and the straightening time associated. Several sources were used to constitute the database, Table I:

- a workload table of giving the estimated and the real value of the work duration (in hours) ; and giving the responsible of the straightening work (the shipyard workers or sub-contractors).
- ships plans to extract the scantling and dimensions of the main frames.
- tables describing all characteristics of sections constituents

This data base used approximately 1000 different entities coming from 15 passenger ships. The DM analysis within the CRISP-DM methodology is divided in four main parts, discussed in the following subsections.

Table I: Description of the database relating to the straightening

Field	Example	Unit	Description
Ship	32	–	Ship identification
Section	v102	–	
R_time_on_surf	100	h/m^2	Time of strengthening work spent by m^2
Section_surf	396,2	m^2	Surface of the section calculated by the multiplication of width and length, it is thus the ground obstruction surface
Section_weight	210,4	t	Total weight of the section (decks constituents + girders + transversal frames + bulkheads + plating..)
Deck_weight	64.2	t	Weight of the deck (plates + stiffeners + girders)
Plate_weight	59.89	t	Weight of deck plates
Delta_stiff	700	mm	Distance between longitudinal stiffeners
Delta_frame	2760	mm	Distance between transversal stiffeners
Thickness	25	mm	Relevant thickness of the deck plates
H_stiff	280	mm	High of longitudinal stiffeners
T_stiff	11	mm	Thickness of longitudinal stiffeners
H_web_frame	1600	mm	High of transversal stiffeners webs
T_web_frame	15	mm	Thickness of transversal stiffeners webs
H_flange_frame	120	mm	Width of transversal stiffeners flanges
T_flange_frame	15	mm	Thickness of transversal stiffeners flanges
Delta_deck	3600	mm	Distance between two deck
Double_bottom	Y	–	Y if double bottom else N
Section_length	17	m	Length of ground obstruction surface
Section_width	31	m	Width of ground obstruction surface
Deck_nbr	4	–	Number of the deck
Slice	4	–	Slice of the ship where the section is (position in the ship axis)
Grade	A	–	Steel grade
Special	N	–	Id special (formed parts of the ships = not precise data about deck surface) thus Y else N
Family	1	–	Clusters of section defined by ship areas
Deck_surf	525	m^2	Surface of the horizontal plates constituting the deck
S_weight_on_surf	0,628	t/m^2	Section weight on deck surface
S_weight_on_length	5,725	t/m	Section weight on section length
D_weight_on_surf	0,192	t/m^2	Deck weight on deck surface
D_weight_on_length	0,1793	t/m^2	Deck weight on section length

4.1 Data description

This step consisted in a presentation of the attributes (fields of the data base), with their distribution and other statistical parameters (minimum, maximum, mean and variance). One of the difficulties which arose during the data base analysis is that most structural attributes show a discrete distribution with one or few dominant modes. For instance, the distance between stiffeners has very often the same value. Those attributes are almost "constant" parameters and thus do not constitute a conclusive information source. In order to minimize this effect, we have replaced some attributes. However, by dividing the weight of plate attribute by the section surface, we obtained information similar to the thickness, but having the advantage to present a distribution much less discrete.

4.2 Data quality

This step listed the problematic recordings (strange distribution, missed values, data in conflict with their physical meaning) in order to take care of them in the next stages. The histograms of all attributes were realized in order to observe the data and to point out the missed or absurd values. We have decided to remove the records corresponding to a straightening time exceeding a threshold value (expressed in h/m^2). Moreover, the data were restricted to the passenger ships.

4.3 Data exploration

This work stage consisted in using different approaches to visualize the correlations existing between

the attributes and the straightening workload in order to finally select the parameters having the most relevant influence on the straightening assessment. In order to fulfil this stage, four different approaches were used:

- a linear correlation analysis trough a dendrograms elaboration
- conventional conditioned distribution histograms
- conventional dots clouds diagrams
- a decision trees analysis

4.3.1 Dendrogram analysis

The dendrogram is the graphical representation of a statistical tool called hierarchical agglomerative clustering. Hierarchical clustering aims to define a sequence of N clusterings of k clusters, for $k \in [1, ..., N]$, so that clusters form a nested sequence. The agglomerative algorithm starts with the initial set of N attributes, considered as N singleton clusters. At each step it proceeds by identifying the two most similar clusters and merging them to form a new cluster. This step is repeated until all attributes are merged together into a single cluster. The similarity among the attributes is measured with the linear correlation coefficient:

$$\rho(x, y) = \frac{\mathrm{cov}(x, y)}{\sigma_x \cdot \sigma_y} \tag{1}$$

$\mathrm{cov}(x, y)$ represents the covariance between x and x. σ_x is the standard deviation of x. This tool is particularly interesting for the analysis of attribute similarities, detecting and eliminating the attributes that are too much linear correlated, or detecting important linear correlations between a goal attribute and input attributes. On a dendrogram, the coefficients displayed represent the minimum linear correlation coefficients between one attribute and a group of attributes or between two groups of attributes. The range of the coefficient is in [0,1].

A dendrogram analysis of the straightening database showed that the interest attribute (straightening in $\mathrm{h/m^2}$) does not seem to be strongly correlated to the other numerical parameters of the database (in particular the scantling variables). This indicates that there are no linear relations between the scantling variables and the straightening cost. However, nonlinear relation may exist.

4.3.2 Conditioned distribution histograms

We used histograms of the straightening workload ($\mathrm{h/m^2}$) conditioned by the thickness plate. Before the use of this method, it is necessary to classify the recordings in clusters following their membership in a particular values range of an attribute; for example, we separate the sections in three groups A, B and C for low, medium or high value of the thickness attribute. We can also visualize the Gaussian curve extrapolated for each cluster and compare the means and standard deviation. The result of this analyze is that the straightening workload grows when the plate thickness decreases. However, results showed a lot of scatter.

4.3.3 Dots clouds diagrams

Dots clouds diagrams were used to display straightening cost as function of the section weight, respectively conditioned by section family and steel grade. The visualization showed a correlation (1/X) between straightening cost and section weight, while we rather expected a similar correlation with deck plate thickness. Each cluster (family and steel grade) points at different places from the curve. We could thus see their influence on the straightening cost.

4.3.4 Decision Tree

The technique of decision trees (DT) is a tool used in classification problems. It aims to elaborate automatic 'if-then' rules. It has a symbolic output and symbolic and/or numerical inputs. The basic

procedure is a search algorithm minimizing a score measure (entropy measure). The implicit goal of this iterative search is to produce a tree, as simple as possible, providing a maximum amount of information about the classification variable of the learning problem. The tree building task is decomposed generally into two subtasks: Tree growing which aims at deriving the tree structure and tests and tree pruning which aims at determining the appropriate complexity of the tree. The main strength of DT is its interpretability, making identification of the relevant attributes for a problem easier. DT is a computationally efficient tool, but less accurate than a neural network.

DT performs a series of tests that permit to separate the data in sub-groups. Typically, if we define a cluster LOW (few straightening work – good quality) and a cluster HIGH (lot of straightening work – poor quality), the tree generates itself in order to dispatch the recordings following their quality. If the separation is well established, the extremity cells tend to have one quality and the path that conducts to these cells tells us the value that the different attributes have to take to obtain little straightening work.

This method provides two interesting outputs:

- the list of discriminatory attributes and theirs relative percentage (weight of section [56.3%], section family [7.7%], thickness [4.7%], etc.)
- the finding of thresholds in order to make the most efficient separation between attributes. For instance, the main branch of the tree is divided with the following rule : "If the section weight $<$ K1 or 163 tons then straightening is HIGH".

We can deduce that section weight is decisive, but other parameters such as family or thickness are also interesting. This method has the major advantage to select the most relevant variables before an analysis by an artificial neural network in order to avoid unnecessary high computing times.

4.4 Data modelling

The visualisation methods presented so far revealed that 'section weight' is the most decisive parameter, followed by 'section family' and 'plate thickness'. None of the scantling attributes ('delta_stiffeners', 'H_stiffeners', 'delta_frame', etc) seem to be correlated to the straightening workload. Because the attribute 'section weight' can fluctuate as function of the section dimensions, we decided to use 'section weight' divided by the section surface. Finally, we selected 'section weight divided by (surface or length)', 'section family' and 'section surface' and 'distance between stiffeners' as input parameters used to establish the straightening estimation formula.

We used the Artificial Neural Network (ANN) technique to find the relation between the five previous attributes and the straightening cost. Artificial Neural Networks, e.g. *Mesbahi (2003)*, have gained popularity for their effective manner to manage complex, multiple input situations and provide a single output. In shipbuilding research, ANNs have been used e.g. for hull resistance prediction, *Couser et al. (2004)*, safety prediction, *Gerigk (2005)*, manoeuvrability prediction, *Ebada and Abdel-Maksoud (2006)*, freight rate prediction, *Bruce and Morgan (2006)*, and propulsion prediction, *Roddy et al. (2006)*. One of their key advantages is their ability to easily model complex, non-linear systems, a feature which is not true of statistical regression methods where an appropriate non-linear function must first be found. An advantage of ANNs over statistical methods is their ability to adapt to new data. Once an ANN architecture has been designed, it can quickly be retrained as new data becomes available. Essentially, a complex set of data can be modelled using ANNs to establish patterns in such systems. The main strength of ANN is its universal approximation capability. It is probably the most accurate data mining method among the available data-driven prediction techniques. Unfortunately, from the point of view of interpretability it is perceived as a black box. ANNs require considerable CPU time during the training stage and may become cumbersome for highly dimensioned input spaces. It is thus advisable to use first methods to that reduce the input space, like decision (or regression) trees or dendrograms.

After having chosen the input parameters, we filtered the input records to obtain the most reliable relation between straightening cost and scantling variables:

- We omitted sections where the straightening work was done by sub-contractors because time measurements (cost table) of straightening are less reliable.
- We considered only the sections having a width near 30 m, because we used the 'weight section divided by the section length'.

This filtering drastically reduced our database.

We used two ANNs: One with 5 input parameters ('family', 'section surface', 'thickness', 'section weight divided by length', 'distance between stiffeners'), a hidden layer of 5 nodes and 1 output parameter ('straightening cost'); One with 9 input parameters ('family', 'section surface', 'thickness', 'section weight divided by length', 'distance between stiffeners', 'steel grade', 'special flag and double bottom flag'), a hidden layer of 5 nodes and 1 output parameter ('straightening cost');

We compared histogram errors of the ANN analysis for the target value ('straightening cost') and the predicted value ('straightening cost as predicted by the ANN') during the validation test of the results. The error dispersion was smaller in the model at 9 inputs than in the model at 5 inputs. Indeed, the correlation obtained was 0.744 for the model at 5 inputs and 0.888 for the model at 9 inputs. However, the method has its limits. Since the recordings were restricted to the works realised by the shipyard workers, the quantity of data exploited was small and thus the robustness of the formula was not very good. When we constructed the error diagrams, we tested the equation on the same data set than the one used to establish the relation. As a consequence, the precision given will be optimistic. We await new data to test the results with a test set different than the learning set.

The analysis of the sensibilities consists in modifying the value of an input attribute (the others being maintained constant) and to observe this impact on the output variable. This methodology has several advantages: visualise a projection of the multi-parameters relation in a two-dimensional plot, understand the validity domains, and outline the influence of different variables. We observed a normal comportment; e.g. the section coming from superstructures have small thickness and are thus characterised by an important straightening contrary to the sections of hull bottom.

5 Conclusions

This data mining study, respecting a rigorous protocol for analyzing the data (since the CRISP-DM methodology was used), has shown how various prediction models of the straightening workload (in h/m^2) for deck plate in passenger ships can be designed. We consider that the data mining analysis process and the related tools available on the market will provide an important added value in modelling the ship industry and we believe that data mining will become an important approach for ship design and production.

Among the various modelling techniques available, we privileged the regression techniques (prediction of a numerical continuous variable) of non-linear type and in particular the hybrid techniques using jointly the decision tree analysis and the artificial neuronal networks. This choice was led by the observation of the very weak linear correlation existing between straightening cost and the scantling variables within this data base.

The advantages of data mining utilisation are:

- Selection of the most relevant variables before artificial neural network (ANN) analysis thanks to the decision tree. The preliminary identification of the key factors improves the effectiveness of the ANN and reduces the computing times.
- Prediction accuracy brought by the ANN method.

References

BRUCE, G.; MORGAN, G. (2006), *Artificial Neural Networks - Application to freight rates*, 5^{th} Int. Conf. Computer and IT Applications in the Maritime Industries, (COMPIT'06), Leiden, pp.146-154

CAPRACE, J.D.; LOSSEAU, N. ; ARCHAMBEAU, D.; BAIR, F.; RIGO, P. (2007), *A data mining analysis applied to a straightening process database*, 6^{th} Int. Conf. Computer and IT Applications in the Maritime Industries, (COMPIT'07), Cortona, pp.186-196, www.compit07.insean.it

COUSER, P.; MASON, A.; MASON, G.; SMITH, C.; KONSKY, B. von (2004), *Artificial neural networks for hull resistance prediction*, 3^{rd} Int. Conf. Computer and IT Applications in the Maritime Industries, (COMPIT'04), Siguenza, pp.391-402

EBADA, A.; ABDEL-MAKSOUD, M. (2006), *Prediction of ship turning manoeuvre using artificial neural networks (ANN)*, 5^{th} Int. Conf. Computer and IT Applications in the Maritime Industries, (COMPIT'06), Leiden, pp.127-145

GERIGK, M. (2005), *Safety assessment of ships in critical conditions using a knowledge-based system for design and neural network system*, 4^{th} Int. Conf. Computer and IT Applications in the Maritime Industries, (COMPIT'05), Hamburg, pp.426-439

MESBAHI, E. (2003), *Artificial neural networks - Fundamentals*, Optimistic Optimization in Marine Design (Eds. Birk and Harries), Mensch&Buch Verlag, Berlin

RODDY, R.F.; HESS, D.E.; FALLER, W.E. (2006), *Neural network predictions of the 4-quadrant Wageningen B-screw series*, 5^{th} Int. Conf. Computer and IT Applications in the Maritime Industries, (COMPIT'06), Leiden, pp.315-335