DE GRUYTER
OPEN

DOI 10.2478/pesd-2014-0018 PESD, VOL. 8, no. 1, 2014

DATA-MINING – A VALUABLE MANAGERIAL TOOL FOR IMPROVING POWER PLANTS EFFICIENCY

Mirela Danubianu[1], Dragos Mircea Danubianu[1], Cristian Teodorescu[2], Lucian Constantin[3]

Keywords: data-mining, mathematical modelling, power production, environmental protection.

Abstract. Energy and environment are top priorities for the EU's Europe 2020 Strategy. Both fields imply complex approaches and consistent investment. The paper presents an alternative to large investments to improve the efficiencies of existing (outdated) power installations: namely the use of data-mining techniques for analysing existing operational data. Data-mining is based upon exhaustive analysis of operational records, inferring high-value information by simply processing records with advanced mathematical / statistical tools. Results can be: assessment of the consistency of measurements, identification of new hardware needed for improving the quality of data, deducing the most efficinet level for operation (internal benchmarking), correlation of consumptions with power/ heat production, of technical parameters with environmental impact, scheduling the optimal maintenance time, fuel stock optimization, simulating scenarios for equipment operation, anticipating periods of maximal stress of equipment, identification of medium and long term trends, planning and decision support for new investment, etc. The paper presents a data mining process carried out at the TERMICA – Suceava power plant. The analysis calls for a multidisciplinary approach, a complex team (experts in power&heat production, mechanics, environmental protection, economists, and last but not least IT experts) and can be carried out with lower expenses than an investment in new equipment. Involvement of top management of the company is essential, being the driving force and motivation source for the data-mining team. The approch presented is self learning as once established, the data-mining analytical, modelling and simulation procedures and associated parameter databases can adjust themselves by absorbing and processing new relevant information and can be used on a long term basis for

[1] Universitatea "Ştefan cel Mare" Suceava, mdanub@eed.usv.ro
[2] INCD-ECOIND – Bucharest, ecoind@incdecoind.ro

monitoring the performance of the installation, certifying the soundness of managerial measures taken and suggesting further adjustments.

1. Data-mining – an overview

The knowledge discovery in databases (KDD) has emerged as a fundamental research area with important applications to science, engineering, medicine, business, and education. It aims to formulate, analyze and implement basic induction processes that facilitate the extraction of meaningful information and knowledge from structured and unstructured data. An important issue regarding this field is the understanding of KDD as a "nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad et. al., 1996). From this point of view, pattern is meant in a very general way. A pattern is whatever an algorithm may find or generate from the data, like a model than scores customers based on a decision tree, on a neural network, or based a regression function, a clustering of the data, or a set of association rules.

The experience of the last years showed that discovering knowledge from huge databases involves a complex process. Over time, several models for KDD process have been proposed. Although are several different between these models, the key message is the same: data mining is just one of several steps in a KDD process. Corresponding to the CRISP-DM model (Wirth and Hipp, 2000), we distinguish the following six tasks:

- business (or problem) understanding - focuses on understanding the project objectives and requirements from a business perspective, and developing initial technical problem definition and a project plan;
- data understanding - based on the results from the business point of view the second step is to get familiar with the available data.
- data preparation - the next step is to construct the dataset where the mining algorithm is to be run on.
- modelling - various modelling techniques are selected, applied, and fine-tuned. In this phase the actual data mining takes place. Based on the identified business goal and the assessment of the available data an appropriate mining algorithm is chosen and run on the prepared data.
- evaluation - at this stage there are good models (from a technical point of view). Here we thoroughly evaluate the model, and review the steps executed to construct the model, to check if we did not miss an important business issue and achieves the desired business objectives
- deployment - can be as simple as generating a report or as complex as implementing a repeatable data mining process.

The analysis of these phases shows us that the KDD process no means only a "push button technology". On the contrary, knowledge discovery is complex, iterative and highly interactive. In each of the above phases it is the analyst as a human being who decides whether to proceed to the next phase, to redo the current phase or even to step back to one of the earlier phases.

Fig. 1 presents these phases and the most important interdepencies between them.

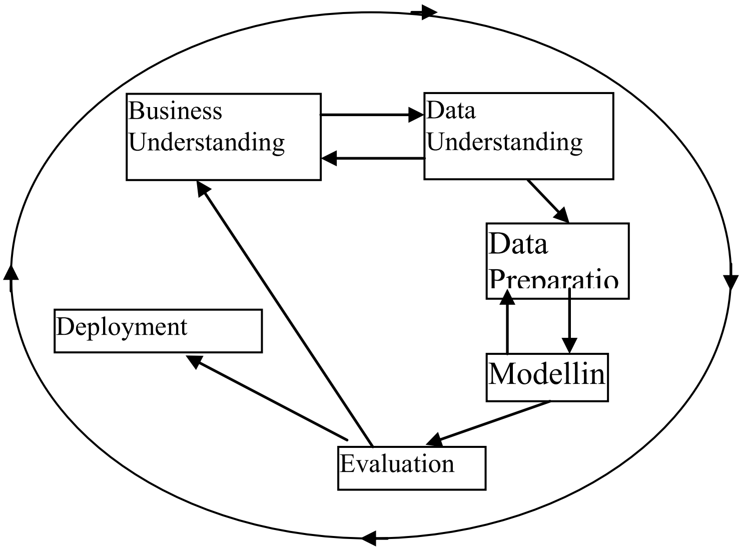

Fig 1. The crisp-DM model for KDD

Data mining involves the application of analysis on large volumes of data using algorithms which produce a particular enumeration of patterns from such data, and it may facilitate the discovery from apparently unrelated data, relationships that can anticipate future problems or might solve the studied problems. On may define data mining as the operation of extracting the interesting and previously unknown information from huge volumes of data.

It is able to solve problems which can be divided into two general categories: prediction and knowledge discovery (or description). Even prediction is the main goal of data mining, often it is preceeded by description. Each of these two problems has some associated methods. For prediction we can use classification or regression while for knowledge discovery we can use deviation detection, clustering, association rules, database segmentation or visualization.

Classification is a supervised learning method which consists in a two-step process. First, by analyzing database tuples described by attributes a model is built. It describes a predetermined set of data classes or concepts. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute. The data set analyzed to build the model form the training data set. In the second step, the model is used for classification. Before that, it is necessary to estimate the predictive accuracy. The accuracy of the model on a data test set is calculated as the percentage of test set sample that are correct classified by the model previously build. To find this percentage, for each test sample, the known class is compared with the model's class prediction for that sample. If the accuracy is acceptable the model can be used for classifying future data tuples for which the class label is not known.

Whereas classification determines the set membership of the samples, the prediction of continuous values can be modeled by *regression.* In this case model design consists on finding a structure for it, on computing optimal value for its parameters and assessing the model quality. The model structure relates the type of mathematical formula that describes the system behavior. Depending on the model structure, regression models may be categorized like: simple linear regression, multiple linear regression, polynomial regression, logistic regression or nonlinear regression. We can also distinguish static and dynamic models. Static models produce outcomes based only on the current input, whereas dynamic models produce outcomes based on the current input and the past history of the model behavior.

Clustering, often referred as unsupervised learning, involve a process that discovers structures in data without any supervision. As the name clustering implies, unsupervised algorithm is capable of discovering structures on its own by exploiting similarities or differences between individual data points on a data set.

Association rules mining is also an important data mining method that aims to find interesting dependencies in large sets of data items. Often these items are stored in transactional databases that must have specific format. This format can be generated by an external process or can be extracted from relational databases or data warehouses. Interesting associations between data items can often lead to information used for decision making.

The algorithms used in data mining are often well-known mathematical algorithms, but in this case they are applied to large volumes of data and to general business problems. The most used are: statistical algorithms, neural networks, decision trees, genetic algorithms, nearest neighbor methods, rule induction and data visualization.

2. Starting the data-mining project

Having the approval and support of the TERMICA power plant top management, a Data-Mining Team (DMT) was established. It included the technical director, the mechanics-energy chief engineer, the chief accounting officer, the heads of the quality and environmental compartments and two junior members that took care of the IT infrastructure and of uploading data in the Project databases. The team included also two experts from outside TERMICA, namely a software professor and an energy and environment consultant, both with experience in data-mining, mathematical and statistical modelling, with good knowledge of the EU BAT (best available techniques) documents related to power production and its environmental impact. The DMT reported to the TERMICA top management on a daily basis so that the Data-Mining Project (DMP) went smoothly and access to all records and data existing at TERMICA was permitted to all DMT members after the signing of a Confidentiality Agreement.

The main objectives for the DMP were established once the team was constituted. These objectives were designed with respect of CRISP-DM steps. There are:

- Establishing the technological, economic and environmental parameters to be analyzed;
- Identifying sources of data, evaluating their coherence and establishing the databases of the DMP
- Assessing the need for supplementary hardware for measuring and monitoring, in order to increase the quality of data; in the first phase, the DMP worked only with existing data, even if their quality could be improved;
- Use of environmental accounting principles and procedures (EMA, 2001; ISO 14051, 2011) for an objective, unbiased analysis of environmental costs;
- Correlation analysis
- Internal Benchmarking (when, how and why have been attained the best operational and environmental performances of the TERMICA power plant)
- Comparison to best available techniques (LCP, 2006)
- Trend identification by regression analysis
- Devising an eco-efficiency index, together with its dynamic evolution, for the TERMICA power plant.

As a working procedure, the TERMICA top management expressly requested that:

- the DMP should use software familiar to local experts, or easy to learn and use by these experts (the Microsoft EXCEL and RapidMiner 5 was subsequently used);
- the DMP procedure and databases will be active, upgradable and at hand even after the Projects ends. Measures were taken to train TERMICA people with the use of the DMP procedures.

Data used in the DMP covered the period 2007-2011.

Existing technical parameters of the installations (specific consumptions, power and heat produced and delivered, excess air to burners, wastewater and ash generated, together with their characteristic parameters, costs and other economic parameters, etc.) were collected and uploaded in an agreed format in the DMP databases.

Assessing the normal (Gaussian) distribution of data was carried out using

- descriptive statistics for each data set (skewness and kurtosis moments of the data distribution were inferred and they are near zero value in almost all cases – indicating a distribution near to normal);
- Histograms and frequency tables showing graphically if the set of data fits the Gaussian case);
- 3 statistical tests specific for testing normality, built up in EXCEL worksheet: Jarque-Bera and Shapiro-Wilk tests. All calculate a *p*-value that is compared to and should be larger than the α-value chosen. All sets of parameter data passed the Jarque-Bera test and most of them the Doornick-Hansen (χ^2) test. Only a few passed the Shapiro-Wilk test.

It was assumed that the data are normally distributed and recommendations were made for installing more performant measuring and monitoring devices to increase in future the quality of acquired operational data.

3. Deriving environmental costs

In order to have a comprehensive image of the technology, economic and environment performance of TERMICA power plant, a detailed environmental cost accounting procedure was developed in order to associate costs to any material and energy flow in the power plant.

Recommendations of UN (EMA, 2001) and ISO (ISO14051, 2011) documents have been used to obtain environmental costs like those exemplified in Table 1.

Table 1. Examples of environmental costs evaluated at TERMICA.

Negative Flux Cost Index (2007=100)	**2007**	**2008**	**2009**	**2010**	**2011**
Electric energy internal consumptions	100	105.6	165.1	114.7	129.0
Thermal Energy not delivered	100	133.5	183.1	135.7	133.3
Ash management costs	100	121.6	221.6	157.1	178.5
Carbon Dioxide associated costs	100	80.4	95.8	88.7	91.4
Cost of flue gases at 140°C	100	109.2	145.3	102.8	102.4
Cost of water / wastewater	100	128.3	144.5	168.2	172.6
Share in total costs (%)	31.1	18.2	19.8	18.9	19.1

Negative flows is the terminology used by ISO 14025 (equivalent to “non-product outputs” of the UN methodology (EMA, 2001) to denominate material and energy flows unintended, that can be reduced but cannot be avoided (losses, by-products, etc.). Table 1 shows an important share of environmental costs associated to power and heat production (up to 31.1% in 2007), yet declining and stabilizing at around 19%.

The structure of these costs, in 2011 is presented in Fig.2 indicating interesting conclusions for the TERMICA management. Examples:

- the larger share of negative costs is due to internal consumptions and losses of electric energy. Reducing these costs means investment in refurbishing existing installations, verifying constantly the state of electric energy consuming, transforming and transporting equipment, good maintenance policies;
- the cost associated to flue gases comes second. In principle, it can be reduced by lowering the gas exhaust temperature (currently 140^{o}C)but this cannot be achieved without risking water condensation and ensueing corrosion problems. Replacing the steel pipes conveying flue gases to stack with pipes made from corrosion resistant materials could be a solution but the investment should follow a detailed cost-benefit analysis;
- costs associated to carbon dioxide emitted can be evaluated once Romania has a market for this substance (CO_2 credits). Indeed TERMICA has managed to reduce costs with CO_2 by trading some of the associated credits allowed to the company. This is why the cost of carbon dioxide emitted has constantly declined to less than 20% of the 2007 value. Not only is this a very favourable environmental issue but TERMICA managed to turn a waste (CO_2) into a source of income.

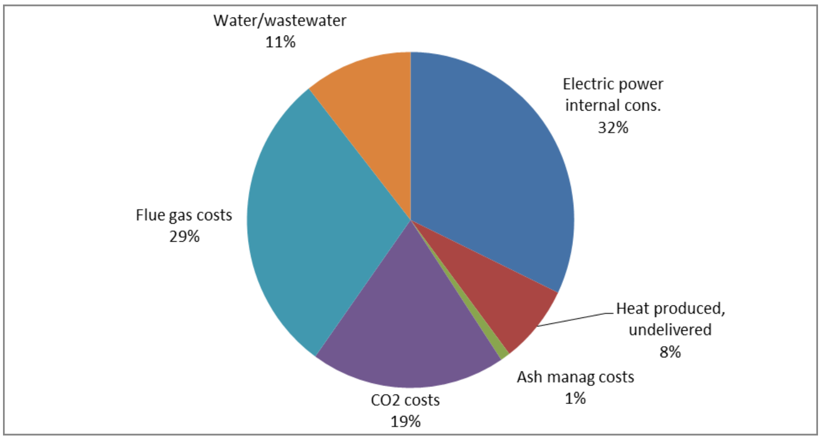

Fig. 2. Structure of negative costs at TERMICA in 2011.

A first glimpse in the operational data recorded at TERMICA has in this way indicated the hot points where managerial action is needed most, presenting objective quantification of spending and taking costs out of overheads.

3. Data correlation

A detailed correlation analysis has been carried out by the DMT with many parameters available in the records kept at TERMICA.

Pearson correlation coefficients were derived and some of them are presented in Table 2.

Some conclusions emerge:

- the ash management costs does not correlate, as expected, with the amount of power / heat produced or with other consumptions (e.g., water). This was explained by the variations of ash content of various solid or liquid fuels used at TERMICA in the period of time analysed;
- the ash and flue gas to stack associated costs correlate best, as expected, as do water and wastewater management costs with the same flue gas costs.

By acting upon, for example, flue gas temperature, practically all other costs (that correlate strongly with this parameter) will also decrease. This points out again on a further detailed analysis of the opportunity of decreasing flue gas temperature.

Table 2. Examples of Pearson Correlation coefficients calculated at TERMICA.

Parameter	**Electric energy internal consumption**	**Heat not delivered**	**Ash Manag.**	**CO_2 emitted**	**Flue gas costs**	**Water / wastewater**
Electric Energy internal consumption	1.00					
Heat not delivered	0.70	1.00				
Ash Manag.	-0.11	-0.01	1.00			
CO_2 emitted	0.36	0.62	-0.76	1.00		
Flue gas costs	0.80	0.78	0.91	0.82	1.00	
Water/wastewater	0.58	0.45	-0.86	0.87	0.88	1.00

4. Internal benchmarking and comparison to BAT

Benchmarking is used in most instances to assess the performances of a given industrial facility to similar installations operated elsewhere (domestic or abroad). The DMP carried out a detailed benchmarking analysis in order to:

- identify the best performances ever attained by the TERMICA installations in the period 2007-2011;
- asking the managers and technical staff to further evaluate how these performances have been attained and what can be done to reproduce them.

Internal benchmarking is, therefore, an approach that shows to operators how well their installations can perform and that ask them to maintain the conditions that enabled such performance. Important improvements can be achieved, in this way, by carefully operating existing equipment, respecting procedures, maintenance periods, etc. Practically no investment is needed but a training and motivation of the personnel.

From the best ever achieved level of performance, the installations should be aligned, by refurbishing and investing in replacing existing, outdated equipment, to BAT reference level of performance (LCP, 2006).

For the case of TERMICA, Table 3 illustrates the dynamics of various parameters in time and their level when compared to BAT documents.

As shown by Table 3, all current performances at TERMICA are well off the level required by best available techniques at hand.

Table 3. Data for internal benchmarking and comparison to BAT.

	BAT	**2007**	**2008**	**2009**	**2010**	**2011**	**%min of max**
Fuel specific consumption for power production	1	1.21	1.16	1.24	1.24	1.20	93.5
Fuel specific consumption for heat generation	1	1.17	1.14	1.22	1.21	1.21	93.4
CO_2/MWh	1	1.78	1.62	1.84	1.79	1.8	88.0
Water m3/MWh	1	1.21	1.14	1.28	1.24	1.21	89.0
Boiler efficiency	1	1.05	1.05	1.05	1.05	1.05	100
Electric energy internal consumption	1	1.21	1.13	1.16	1.17	1.18	93.3

5. Regression analysis and trend identification

Data from the 2007-2011 period can be used to identify current trends in TERMICA performances and evaluates the period of time until best available technology efficiencies are attained (Fig.3).

Fig. 3 indicates that continuing the current operational set-up and efforts to improve at the current pace (small investments in burning hardware, in water and ash management, etc.), TERMICA will reach the BAT level for water consumption in 2017, the BAT level for specific fuel consumption per 1 MWH in 2022 and the same consumption per 1 Gkal in 2040! All these identified trends are correct (as they indicate a progress towards BAT level) but the period of attaining this BAT level is far too long. The level of electric energy internal consumption is increasing instead of decreasing and aligning to the BAT value (here represented by unity). This is a hot point that must be addressed with priority by managers and engineers.

While the lag for attaining BAT consumptions in the case of water maybe acceptable, the other trends are barely so and the management must take decisions to refurbish the equipment and/or change the technology in order to align to BAT as soon as possible.

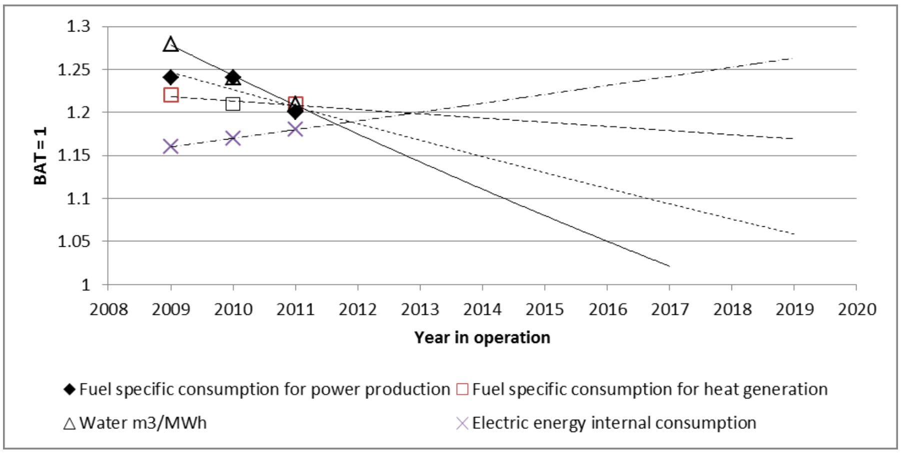

Fig. 3. Trends in TERMICA operational parameters based upon current set-up.

Conclusions

The paper presented an overview of the data-mining analysis that exploit the vast amount of data recorded by industrial installations during their years of operation in view of identifying hot points, trends, best efficiencies ever attained, etc.

With a dedicated team of engineers, economists, environmentalists, external consultants and having the full support of the top management at TERMICA power plant in Suceava (Romania, 450km N of Bucharest), the data mining-project carried out by the team led to the following findings:

a. there is a good amount of relevant data at TERMICA that allow a sound data-mining analysis; further updating of the created databases, in the format agreed during the Project will enable the data-mining procedure to absorb extra information and produce updated conclusions aout TERMICA process efficiencies and environmental footprint;

b. procedures for testing the coherence of recorded data, evaluate their Gaussian distribution were implemented in EXCEL and recommendations for further improve the quality of data by investing in new monitoring equipment for data acquisition were made to the Company management;

c. the Project Team helped local specialists to derive more objective estimates for their environmental costs by implementing procedures recommended by UN and ISO standards; it showed that these environmental costs amount to 18-31% of total Company expenses, in the period of time analysed (2007-2011);

d. the structure of these costs showed that the hot points, where costs are the higher, are the internal consumption of electric power generated and the costs associated to flue gas. The environmental cost accounting procedure should remain in place as it ascertain best what the environmental foot print of TERMICA is and how it evaluates. Managers should reduce this footprint by acting where the environmental costs are the;

e. correlation analysis enabled the parameters that correlates best. Acting upon one of them will result in a corresponding improvement in all other correlated parameters;

f. a benchmarking analysis was carried out having at hand the operational data included in the data mining project. The external, classical, benchmarking was carried out using best available techniques documents as reference. The internal benchmarking indicated the year 2008 as the best on in the recent history of the power plant. By simply reproducing the conditions valid in 2008, TERMICA could improve by 6-12% level of efficiency, without any investment;

g. regression analysis pointed out that, with the notable exception of internal electric energy consumptions, all other important operational parameters tend to approach the levels specified by the best available techniques documents, though the period in which those levels will be attained is considerable. This constitutes an objective assessment of the current and long-term performances at TERMICA and will help managers to decide where to act in the first place in order to improve. First the must reverse the increasing tendency in internal electric energy consumption, Then consumptions of fuel for 1MWh and 1 Gkal must be addressed.

The Project showed the importance of using every source of information and treats it using state-of-the-art methods in order to get the most out of such information. It should remain as a basic tool for local managers and specialists.

References

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996), *The KDD process for extracting useful knowledge from volumes of data*. Communications of the ACM, 39(11):27-34

LCP (2006), *Reference document on best available techniques for large combustion plants,* available at http://eippcb.jrc.ec.europa.eu/reference/lcp.html , accessed August 30, 2013.

EE (2009), *Reference document on best available techniques for energy efficiency* available at http://eippcb.jrc.ec.europa.eu/reference/ene.html accessed August 30, 2013.

ECMA (2006), *Reference document on economic and cross-media effects*, available at http://eippcb.jrc.ec.europa.eu/reference/BREF/ecm_bref_0706.pdf . accessed August 30, 2013.

Ruan, Da et al, (eds.) (2005), *Intelligent Data Mining,* Springer Verlag, Berlin

EMA (2001), *Environmental and management accounting*, UN Division for sustainable development, UN, New-York, 2001

ISO 14051 (2011), *Environmental Accounting – Material Flow Cost Accounting – General Framework* , available at , http://www.iso.org/iso/catalogue_detail?csnumber=50986 accessed August 30, 2013

Nisbet, M. et al (2012), *Demolition and Deconstruction: Review of the Current Status of Reuse and Recycling of Building Materials*. Available at ftp://ftp.tech-env.com/pub/Retrofit/AWMA%20paper_WM1b.pdf , accessed April 3, 2012

SYMONDS (1999), *Construction and Demolition Waste Management Practices and Their Economic Impacts*. A Study Prepared by SYMONDS Group (UK) in association with ARGUS (Germany), COWI Consulting Engineers and Planners (Denmark) and PRC Bouwcentrum (the Netherlands) for The European Commission, DG-XI.

Venta, G. and Nisbet, M. (2000), *Potential for Reduction of CO2 Emissions in Canada through Greater Use of Fly Ash in Concrete*, CANMET/ACI International Symposium on Concrete Technology for Sustainable Development, Delta, Canada.

Voina, N.I. (1981), *The Theory and Practice of Pulverized Fly Ash Utilization* (in Romanian), Ed. Tehnica, Bucharest, 488pp.

Wirth, R. and Hipp, J., (2000), *CRISP-DM: Towards a standard process model for data mining.* In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, pages 29-39, Manchester, UK.