Paper 161-2007

Methodology for Railway Demand Forecasting Using Data Mining

Giovanni Melo Carvalho Viglioni, Instituto Militar de Engenharia, Rio de Janeiro, Brasil

ABSTRACT

After the organization to resolve its operational problems, comes the necessity for systems to support decision making. The research area data mining grows quickly to take care of these new necessities. However, the use of data mining techniques becomes difficult due to the lack of a complete and systematic methodology for the knowledge discovery in database. This dissertation presents a model of the formal process of development of systems of discovery of knowledge in database for the prediction of railroad demand, that includes a systematic and rigorous methodology, which integrates the methodologies: CRISP-DM, SEMMA, FAYYAD, and an interactive environment for the implementation of these systems. The methodology proposal integrates the cited methodologies and was applied in a customer transport request database of MRS Logística, during the period of Dec, 1st of 2003 until Oct, 31st of 2006. This application is main objective was to validate the methodology proposal according to the criteria of the respective company. The conclusions of the case studies allowed us to show the relevance of the MPDF-DM methodology in the forecast of railroad demand.

INTRODUCTION

The efficient management of any company requires planning, whether it is public or private, industrial, the retail sector, or services. In order to be effective, it is necessary to have expectations of the future conditions under which the company will operate and of as if they relate the elements this expectation.

The manager of a railroad, in order to make the right decisions, must know the expectation of transport growth in order to put in place the necessary equipment and the man power, and also what the main factors are that affect this demand and the supply capacity of the arrival and departure terminals. Both the strategic and operational decisions of a company require the exploration of the current relationship between the elements that compose the reality in which the company exists. To support the corporate decisions cited above, the companies look to create systems and procedures, in order to explore scenarios based on quantitative and/or qualitative information.

Around the world, railways are an important means of passenger and cargo transportation. Railway transportation is characterized by its capacity to carry great volumes with high energy efficiency, mainly in cases of medium-to-long distance transport. It is also safer than road transport, with lower incidence of accidents and robberies/theft. FIG. 1 presents the distribution of cargo transport in Brazil, where 24% corresponds to railway transport. (ANTT, 2006).

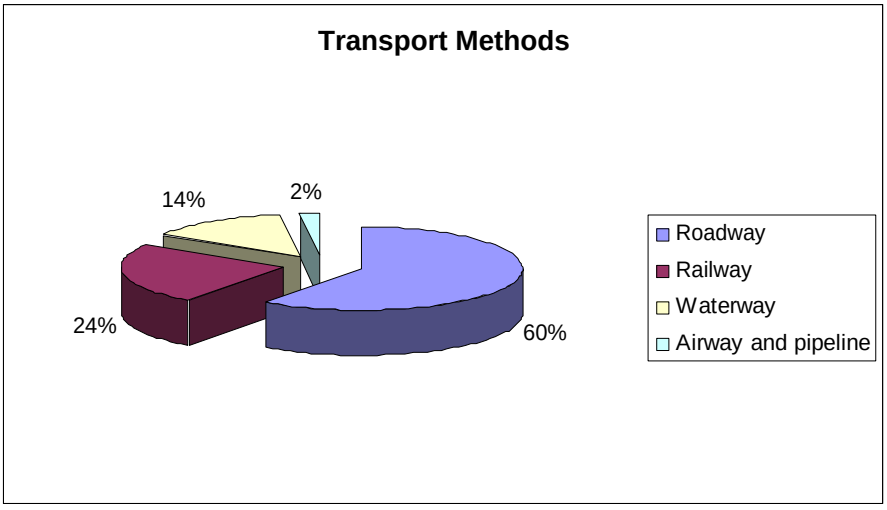

FIG. 1: Transport methods in Brazil – 2005.

Developing countries, such as Brazil, need to improve the social-economic indices of the population, having the necessity of greater commercial exchange of merchandise with other countries. There is a consensus that railway and waterway are the best means of transport since trains carry merchandise to the port, and ships complete the trip to the final destination. As shown in TAB. 1, there is an unexplored margin of railway growth in Brazil. (MELO and

MEZZONATO, 2005.)

TAB. 1: Cargo Transport Methods (%) – 2004.

Methods / Country	Waterway	Railway	Roadway
Australia	4	43	53
Brazil	14	24	62
Canada	11	46	43
China	13	37	50
USA	25	43	32
Russia	11	81	8

The general objective of this work is to develop a methodology capable of identifying useful relationship standards in the demand forecasts, in order to assist in the sizing of railway transport. These forecasts must encompass a great deal of the short-term decisions faced daily by the railway manager, to a deep level (in detail in regards to products) and give important assistance to investment decision-making when it comes to traffic, transport, and terminal capacity, in the long-run.

OVERVIEW OF RAILROAD DEMAND IN THE BRAZIL

According to SILVEIRA (2003), at the beginning of the 17th century, transportation in Brazil extended from the coast to the interior and vice-versa, using primitive methods, that is, those which do not make use of mechanical traction, but of power provided by man and beast, water currents, and gravity and wind power. In land transport, products were loaded on the backs of men, by dragging, on sleds, and wheeled-vehicles. With the arrival of the railroads, which were the result of the expansion of the revolution in transportation, this method of transport began to serve the export of coffee, cultivated in the massive plantations of the southeast. Therefore, railways strengthened the export of agricultural products.

Railways that were created to serve a specific demand are presented below.

TEREZA CRISTINA RAILWAY (FTC)

Inaugurated in 1880 to serve the demand of coal transportation in southern Santa Catarina state.

CARAJÁS RAILROAD (EFC)

Constructed to serve the demand of transporting iron-ore between Carajás in southern Pará state, and the Port of São Luis, in the state of Maranhão, it was inaugurated in 1985.

VITÓRIA-MINAS RAILROAD (EFVM)

905 kilometers in length, EFVM is a subsidiary of the Vale do Rio Doce Company (CVRD) their objective being the transport of iron-ore from Minas Gerais for export at the Port of Tubarão.

FERRONORTE RAILWAY

The Ferronorte railway was designed mainly to serve the demand of the Itamarati Group, which was the main exporter of soy in Brazil during the 1980's. It is currently 504 kilometers long and is not yet completed.

NORTH-SOUTH RAILWAY

The initial design of the North-South Railway projected the construction of 1,550 kilometers of track, passing through the states of Maranhão, Tocantins, and Goiás. It was designed to serve the demand of soy transportation between the mid-western region and the northern ports.

TRANS-NORTHEASTERN RAILWAY

Planned over 100 years ago, the railway that crosses the Brazilian outback finally got some tracks in 1990, but its construction was halted due to lack of government funds in December of 1992 according to DNIT (2006). When the project was resumed, it became known as "Nova Transnordestina" (New Trans-Northeastern) and it's objective is to serve the demand in the northeastern region, connecting the ports of Pecém in the state of Ceará and Suape in the state of Pernambuco.

JARI RAILROAD(EFJ)

The Jari Railroad, built to transport timber that supplies the cellulose factory of the Jari Project, began operation in 1979, according to SANT' ANNA (1998). It is located in the north of the state of Pará, near the border with the state of Amapá.

AMAPÁ RAILROAD (EFA)

Inaugurated around the end of September in 1956, it is dedicated to the transport of manganese-ore from the

deposits of the Serra do Navio mountains, to the Port of Santana, located near the city of Macapá on the left bank of the northern canal of the Amazon River.

TROMBETAS RAILROAD (EFT)

Constructed to carry bauxite from the mine to the Port of Trumbetas in the state of Pará. It was inaugurated in 1979.

"STEEL" RAILWAY

In the early 1970's, a preliminary study was done for the establishment of a modern railway connection between Belo Horizonte and São Paulo. The results of the study were published with a great deal of enthusiasm by the press in May of 1973, given the name "Steel Railway," its objective being to serve the demand of ore transportation. In spite of its enormous dimensions was still justifiable, taking into consideration the notable economic performance of Brazil in the early 1970's, the famous age of the "Brazilian Miracle." The economy had grown at levels of more than 10% annually between 1968 and 1974, and people did not expect the growth rate to drop below 8% until at least 1980. After many crises which affected the construction of the railway, on April 14, 1989, the "Steel Railway" was concluded, finally allowing rail traffic after 14 years of work. The so-called "Thousand-Day-Railway" had actually become the "Five-Thousand-Ninety-Eight-Day-Railway".

DEMAND FORECASTING

The planning and control of transport activities depend on accurate estimates of the volume of services to be provided by the company. Such estimates are typically made by planning and forecasting, according to BALLOU (2006). Such techniques can be divided into two main groups of approach: quantitative and qualitative, according to MAKRIDAKIS, *et al.* (1998).

QUALITATIVE

The qualitative forecasting techniques, also called "subjective" or "criteria judgment based" techniques, are those that make use, primarily, of the human capacity to generalize and extrapolate.

QUANTITATIVE

Quantitative forecasting techniques are those that use historical data to mathematically calculate extrapolations of future data. According to MAKRIDAKIS, et al. (1998), forecasting, using quantitative techniques, can be applied when:

1. When past information is available;
2. The information can be quantified in mathematical terms;
3. It would be possible to assume that some aspects of the standard verified in the past will continue in the future. This statement is also called "estimation of continuity".

DATA MINING

Constant advancements in the area of Information Technology have made the storage of multiple and very large databases possible. Technologies such as the Internet, database management systems, bar-code readers, less costly and larger secondary memory devices, and information systems in general are some examples of resources which have made possible the proliferation of countless databases of scientific, government, administrative, and commercial nature.

It is not feasible for people to analyze great amounts of data without the assistance of appropriate computational tools. Therefore, the development of tools of an automatic and intelligent nature becomes essential for analyzing, interpreting, and correlating data in order to develop and select strategies in the context of each application.

To serve this new context, the area of Knowledge Discovery in Databases (KDD), came into existence with great interest within the scientific, industrial, and commercial communities. The popular expression "Data Mining" is actually one of the stages of the Discovery of Knowledge in Databases. Both will be detailed below.

The term "KDD" was formally recognized in 1989 in reference to the broad concept of procuring knowledge from databases. One of the most popular definitions was proposed in 1996 by a group of researches. According to FAYYAD, et al. (1996): "KDD is a process with many stages, non-trivial, interactive, and iterative, for the identification of comprehensible, valid, and potentially useful standards from large data sets".

DISCOVERY ASSOCIATIONS

The classic task of searching for association rules (also called associative rules) was introduced by AGRAWAL, *et al.* (1993). Intuitively, this task consists of finding sets of items that occur simultaneously and frequently in a database.

DISCOVERY OF SEQUENCES

This is an extension of the task of discovering associations, which takes into account the temporal aspect between the transactions registered in the database.

CLASSIFICATION

One of the most important and popular KDD tasks is that of classification. This task can be understood as the search

for a function which allows the correct association of each Xi register of a database to a single categorical label, Yj, called “class.” Once identified, this function can be applied to new registers in order predict the class into which they fit.

SUMMATION

Summation, also called “description of concepts,” consists of identifying and presenting the main characteristics of the data contained in a data set in a concise and comprehensible way.

CLUSTERING

“Clustering”, also known as “Grouping,” is used to partition the registers of a database into sub-groups or “clusters,” in a way that the elements of a cluster share a set of common properties which distinguish them from the elements of other clusters. The object of this task is to maximize intra-cluster similarity and minimize inter-cluster similarity. Unlike classifications having predefined labels, it is necessary for clustering to automatically identify the labels and this is why it is also known as “non-supervised induction”.

TIMES-SERIES FORECASTING

According to CARVALHO (2005), detecting regularities in phenomena that occur throughout time and the ability to predict future trends are some of the most important tasks in today’s world. Forecasting temporal-series, such as future free-market prices, stock exchange trends, and patient prognostics are not easy activities. This is because the number of parameters involved are great and the discovery of cycles or repetitious patterns do not always take a clear form because mathematical techniques have their limits when faced with nonlinear, dynamic phenomena, like chaos.

DEVIATION DETECTION

The object of deviation detection is to identify changes in previously perceived patterns. Its usage has grown significantly in recent years, being used extensively to detect fraud in credit card usage, health insurance plans, tax collection, locomotive fuel consumption, among others. This task has as its goal, finding data sets that don’t follow the correct behavior or design. Once found, they can be treated, or discarded for use in the KDD process. This makes the data evaluation important in the sense of discovering growing probabilities of deviation, or risks associated to the many objectives initially planned in data mining.

PROPOSED METHODOLOGY

The proposed methodology is called “Methodology for Railway Demand Forecasting Using Data Mining, (“Metodologia para Previsão de Demanda Ferroviária utilzando Data Mining,” or MPDF-DM, in Portuguese.). Taking into consideration the inherent complexity normal in the processes of Knowledge Discovery in Databases, this methodology is based on principles of activity planning. Thus, according to the objectives of each KDD application, the steps in the knowledge discovery process are planned before being executed. The application of the proposed KDD methodology is divided into four stages as shown in FIG.2.

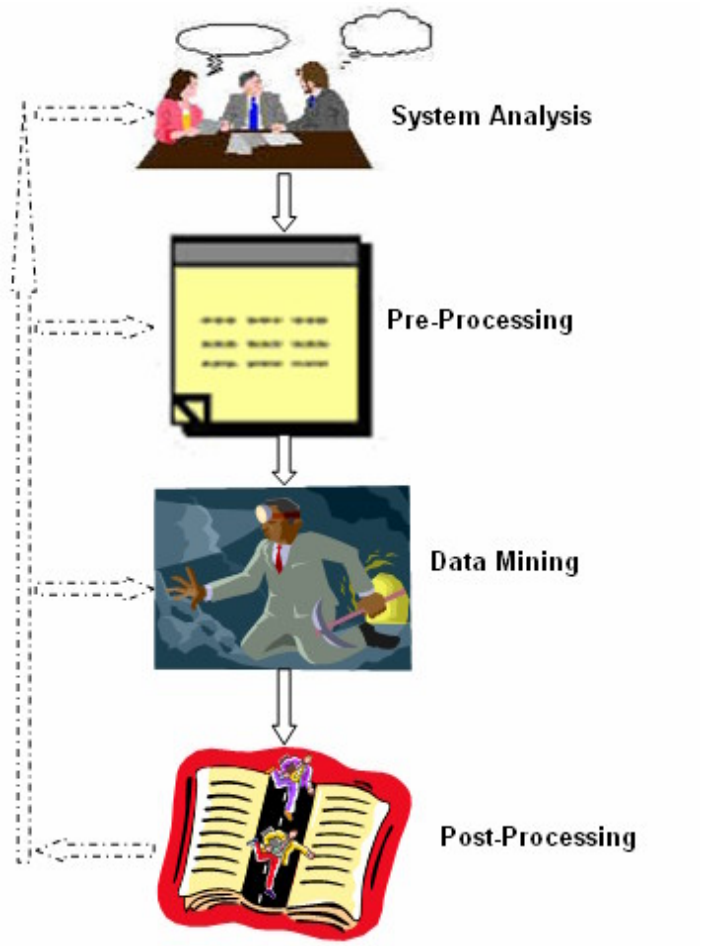

FIG. 2: MPDF-DM Methodology.

The proposed methodology suggests an interactive and iterative process, in which, depending on the results obtained, the KDD analysts can return to any step previously taken in search of better results. In order to do so, the methodology requires detailed documentation of the actions carried out and the results obtained. For this purpose, documentation models are recommended which are supported by a basic line of reasoning, to assist in the choice of procedures to be adopted in light of the diversity of situations and possibilities -- also allowing for documentation of the entire project.

SYSTEM ANALYSIS

This, the first stage of the methodology, has as its main objective to define the types of inquiries to be carried out with the application of the techniques of the KDD process. It is in this stage that the interested people are identified, as well as what is desired to be solved or improved, and the objectives of the process including their respective improvements, as in FIG. 3.

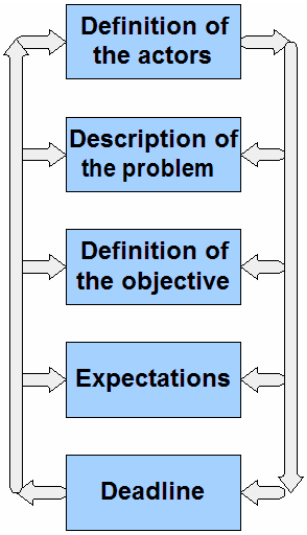

FIG. 3: Activities of the System Analysis Stage.

PRE-PROCESSING

This stage encompasses the functions related to capturing, organizing, treating and preparing the data for the Data Mining stage, having a fundamental importance in the knowledge discovery process. It includes the areas of correction of incorrect data, to the adjustment of data formatting, to the algorithms of Data Mining to be used. It is divided into up to six activities, as shown in FIG. 4.

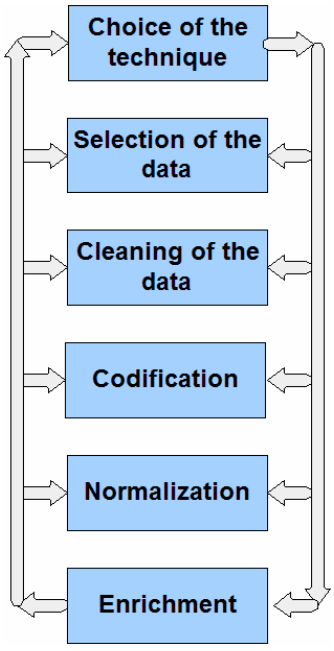

FIG. 4: Activities of the Pre-Processing Stage.

Choice of the Technique

Once the desired results of the KDD process are defined, it's time to choose which techniques can be used that are more appropriate for obtaining results with greater precision.

Selection of the Data

The activity of selection is mandatory in the stage of pre-processing, due to the necessity to report the origin of the information in the KDD process, whether it be transactional or data-warehouse.

Cleaning of the Data

Cleaning the data only makes sense when absent information, inconsistencies, and values not pertaining the domain are found in the process; therefore being an optional activity. When the origin of the information is a data-warehouse, the possibility and necessity of cleaning diminishes, since one of the processes in the creation of a data-warehouse is cleaning the database, according to KIMBALL and ROSS (2002). If the origin is a transactional bank, the possibility increases, according to BRAGA (2005).

Codification

This is the pre-processing activity responsible for the way in which the data will be represented during the KDD process.

Normalization

Normalization consists of attributing a new range to an attribute in such a way that the values could be within the new range in a specified interval, such as from -1.0 to 1.0 or from 0.0 to 1.0. Such an adjustment becomes necessary to prevent some attributes from presenting a larger range of values than others, thereby influencing the tendencies of some methods of Data Mining.

Enrichment

Enrichment consists of the ability to add more information to the existing registers so that these supply more elements to the knowledge discovery process.

DATA MINING

Data Mining is the main stage of the proposed methodology wherein occurs the effective search for new and useful knowledge from the data. Because of this, Data Mining and the KDD process are referred to in an indistinct way, as though they were synonymous, by many authors such as: FAYYAD, et al., (1996), GOLDSCHMIDT and PASSOS (2005), CARVALHO (2005), and BRAGA (2005). The activities of this stage are represented in FIG. 5.

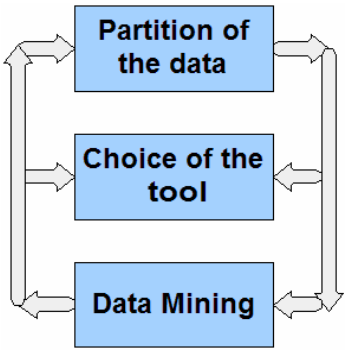

FIG. 5: Activities of the Data Mining Stage.

POST-PROCESSING

This stage involves the simplification and presentation of knowledge models generated by the Data Mining stage. In general, it is in this stage that the KDD specialist and the domain application specialist evaluate the results obtained and define new alternatives of data inquiry. The last stage of the methodology has the activities of FIG. 6.

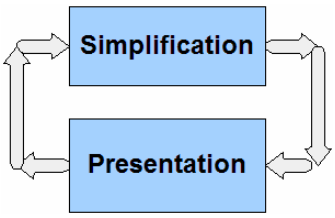

FIG. 6: Activities of the Post-Processing Stage.

CASE STUDY

The goal of this chapter is to present the application of the proposed methodology, for evaluation of the performance of transportation demand planning by the clientele of MRS Logística, S.A., which is a company located in Juiz de Fora, in the state of Minas Gerais, Brazil, that provided all the necessary data so that the methodology could be tested.

APLICATION OF METHODOLOGY

MRS Logística provided the data related to the request of Gross Tonnes (a unit of measurement used in rail transport) on behalf of their customers. This base with a total of 562,746 registers, covers the period between December 1, 2003 and October 31, 2006 and the information is contained in TAB.2.

TAB. 2: Data Structure of MRS Logística.

Column	Description
NU_FLUXO_TRPT	Number of the transport flow
DT_DEMD_VAGAO	Date of the demand of car
QT_VAGAO_SLTD	Amount of requested cars for the transport
PS_TU_SLTD	Weight in requested TU for the transport
DC_MERC_RSMD	Commodity to be carried
NM_ABRV_CLIE	Abbreviated name of the client
SG_PATIO_FRVR_ORIG	Acronym of the origin marshalling yard
NM_PATIO_FRVR_ORIG	Name of the origin marshalling yard
SG_PATIO_FRVR_DEST	Abbreviature of the destination marshalling yard
NM_PATIO_FRVR_DEST	Name of the destination marshalling yard
DC_PROD	Description of the product
SG_TERM_CLIE	Acronym of the Client's Destination Terminal
NM_TERM_FRVR_CLIE	Name of the Client's Destination Terminal

Since the methodology is iterative, some activities were repeated because different techniques were used. The first activity was to identify the actors involved in the KDD process, followed by a description of the problem faced by the company in the demand-forecast area, that being, the difficulty of making the production programming of the company in advance. The next activity was to define the objective, which was improving the predictability of the company's programming production process. The expectation of the model generated by the process was defined as being any model that generated a maximum 10% margin of error according to the Mean Absolute Percentage Error metric (MAPE). The deadline for the execution of the entire process was November 30, 2006.

Due to the fact that the model generated was a day in advance, the chosen techniques are short term, according to BALLOU (2006), Moving Average, Exponential Smoothing, Exponential Smoothing with Trend (Holt's Method), Linear regression, Fuzzy Logic, Neuro-Fuzzy, and Neural Networks Artificial (NNA).

The activities of Data Selection and Data Mining were executed for the forecasting techniques of Moving Average, Exponential Smoothing, Holt's Method, Linear Regression and Fuzzy Logic. All these techniques were adjusted in order to obtain a lower MAPE. While mining, the differences which appeared in the filling in of the form were those of Moving Average, Exponential Smoothing, Holt's Method, and Linear Regression for which the tool used was Microsoft Excel 2003 and also Fuzzy Logic, for which Fuzzy Rules 2001 was used.

For the Neuro-Fuzzy technique, additional activities were carried out in addition to the ones carried out previously. In the activity of Data Selection, Weights from one and two days prior were selected and for the activity of Normalization it was necessary to normalize the weights in a way that the performance of the technique increased. The software used for this technique was Matlab with the ANFIS package.

The last technique used was the Neural Network Artificial (NNA), utilizing the software SAS Enterprise Miner that used all the information described in TAB. 2., with the necessary modifications for optimum performance of the technique being added to the information regarding the average daily value of the dollar and daily import/export trade values, by means of the Enrichment and Normalization activities. It was necessary to the method to accomplish the Codification activity in order to accommodate the attributes of the client's name as well as the origin and destination platform acronyms. Other activities carried out were Data Selection and Data Mining.

After executing the seven techniques, the last stage of the MPDF-DM methodology was fulfilled, that of Post-Processing. When the winning technique was chosen by means of the MAPE metric, as defined in the Expectation of the Model activity, it obtained the results shown in TAB.3.

TAB. 3: Results of Forecasting Techniques.

Technique	MAPE
Exponential Smoothing	3,87
Fuzzy Logic	4,05
Holt's Method	3,86
Linear Regression	10,09
Moving Average	3,95
Neural Network	2,17
Neuro-Fuzzy	4,60

Seven forecasting techniques were used for railway transport demand requests. Among them, the technique using Neural Nets received superior results compared to the others based on what was defined in the Expectation of the Model of Knowledge activities, therefore it was the chosen method.

Since the objective of this case study is not the performance comparison of the techniques, but the usage of the methodology, there were variations in the input variables of the models, in such a way that each technique got the best performance, therefore we can not say that one technique is better than another.

It is important to highlight that from all the MPDF-DM stages, the Pre-Processing stage was the most time-consuming at approximately 70%, due to the fact that it was where the entire process of data selection and preparation occurred, independent of the technique used.

CONCLUSION

The theoretical research carried out showed that in spite of the existence of methodologies in the area of Discovery of Knowledge in Database, difficulties are still found in the application of the process. This is generally due to indeterminate characteristics of these systems and the lack of a specific, complete, standardized, methodology for the development of these projects in order to guarantee the attainment of trustworthy and quality systems.

Being based on classic methodologies, the initiatives of development and the application of knowledge discovery systems are not standardized. In order to overcome these difficulties, this work has proposed a specific standardization model for the task of forecasting railway demand in the KDD process. This model encompasses a rigorous and systematic methodology for the projects called MPDF-DM which combine the standardization of the CRISP-DM methodology, with its forms, along with the stages of Fayyad's methodology.

With the finalization of the case study, it can be concluded that the MPDF-DM methodology, followed step by step, leads to the efficient development of the KDD process. The case study also served to show that obtaining trustworthy and quality projects can be guaranteed with the inclusion of formal methods and the usage of forms in the development process.

The use of SAS Enterprise Miner, specialized software for use in Data Mining, made the entire process easier because it reduced time in the Data Mining stage, due to its use of the Neural Net technique. With other techniques, it was necessary to develop the software, thus delaying the mining stage. This extra time can be used to develop new models which benefit the organization.

The conclusions of the case studies showed the relevance of the MPDF-DM methodology in the attainment of Data Mining results, all coming from hypotheses raised by users and searching, step by step, for ways to prove or disprove these hypotheses.

REFERENCES

Agrawal, R.; Imislinski, T; Swami, A. Mining Association Rules Between Sets of Itens in Large Databases. ACM SIGMOD Conference Management of Data, 1993.

ANTT, Agência Nacional de Transportes Terrestres. Evolução Recente do Transporte Ferroviário. Available: www.antt.gov.br/concessaofer/EvolucaoFerroviaria20060614.pdf [captured in 15/07/2006], 2006.

Ballou, Ronald H. Business Logistics/ Supply Chain Management. Prentice Hall, 2004.

Braga, Luis Paulo Vieira. Introdução à Mineração de Dados. Rio de Janeiro, E-papers, 2005.

Carvalho, Luís Alfredo Vidal de. Data Mining: A Mineração de Dados no Markenting, Medicina, Economia, Engenharia e Administração. Rio de Janeiro, editora Ciência Moderna, 2005.

DNIT, Departamento Nacional de Infra-estrutura Terrestre. DNIT recomenda Transnordestina: Agora Vai. Available: http://www.dnit.gov.br/noticias/Transnordestina/view?searchterm=transnordestina [captured in 02/11/2006], 2006.

Fayyad, Usama M.; Piatetsky-Shapiro, Gregory; Smyth, Padhraic; Uthurusamy, Ramasamy. Advances in Knowledge Discovery and Data Mining. Massachusetts, USA, editora The MIT Press, 1996.

Goldschmidt, Ronaldo; Passos, Emmanuel. Data Mining – Um Guia Prático. Rio de Janeiro, editora Campus, 2005.

Kimball, Ralph, Ross, Margy. The Data Warehouse Toolkit. Guia Completo para Modelagem Dimensional. Editora Campus. Rio de Janeiro, 2002.

Makridakis, Spyros; Wheelwright, Steven C.; Hyndman, Rob J. Forecasting: Methods and Application. John Wiley & Sons, New York, 3a. ed, 1998.

Melo, L.; Mezzonato. Ferrovias: Integração e Crescimento Econômico. Seminário: Ferrovias – Integração e Crescimento Econômico, São Paulo, p. 12 – 13, 2005.

Sant'anna, José Alex. Rede Básica de Transportes na Amazônia. IPEA – Instituto de Pesquisa Econômica Aplicada. Brasíla, 1998.

Silveira, Márcio Rogério. A importância Geoeconômica das Estradas de Ferro no Brasil. Tese (Doutorado em Geografia) – Universidade Estadual Paulista, Presidente Prudente, São Paulo, 2003.

ACKNOWLEDGMENTS

To SAS Brazil, for making their software available and especially to Andrea Szyfer for her attention and dedication.

CONTACT INFORMATION

Giovanni Melo Carvalho Viglioni
Institutuo Militar de Engernharia
Rua Prof. Coelho e Souza, 28
Juiz de Fora, Minas Gerais – 36016-110 - Brazil
+55 (32) 8845-6974
viglioni@terra.com.br

Marcus Vinícius Quintella Cury
Institutuo Militar de Engernharia
Praça General Tibúrcio, 80
Rio de Janeiro, Rio de Janeiro – 22290-270 - Brazil
+55 (21) 2546-7028
mvqc@uol.com.br

Paulo Afonso Lopes da Silva
Institutuo Militar de Engernharia
Praça General Tibúrcio, 80
Rio de Janeiro, Rio de Janeiro – 22290-270 - Brazil
+55 (21) 3820-4120
estatística@estatistica.eng.br