

APPLICATION OF DATA MINING PROCESS TO EXTRACT STRATEGIC INFORMATION ON TRANSPORTATION PLANNING

Wanderley Freitas ¹, Yaeko Yamashita ², Willer Carvalho ³

^{1,2,3}University of Brasilia, Brasilia, Brazil

Email for correspondence: wanderley.ppgt@gmail.com

ABSTRACT

Decision support systems provide analytical models and tools to analyze large amount of data. The Transportation System's database, specifically, stores voluminous data gamma that are necessary to planning, managing and controlling. Due brief time period to examine and interpret the data, data are not explored in its max potential. The data mining process bring forth significant patterns and rules to support managers on decision-making activities during the transportation planning process. This methodology was applied in a research of rural scholar transportation service database. The research was held in 2.277 Brazilian municipalities and developed by CEFTRU/FNDE (2007a, 2007b). As a result, the data mining process admitted 45 associative rules whose measure of interest input parameters values are support $\geq 30\%$, reliability $\geq 30\%$ and lift ≥ 1.10 .

1. INTRODUCTION

The transportation system is a set of elements: actors, organized and interrelated activities that are mutually influenced. Its objective is allowing the indispensable displacement of people and goods provided by a complex roadway, vehicles, terminals and users. (CEFTRU/FNDE, 2008a, 2008b). For transportation studies, the information about the entire planning process is necessary. The information system consists in storage, recovery, filtering, monitoring and rearrangement on demand.

The great amount of data produced by the transportation system is a coefficient that damages timely the quality, quantity and efficiency interpretation for the effectiveness of the result to be reached.

From this vision, a new generation of methods and techniques capable to help managers and decision-makers executives to seek for useful transportation knowledge into the

database is necessary to be created. According to Fayyad (et al., 1996), the data mining process – DM emerged as one of the main support solution in the knowledge seeking process of a transportation database. Thus, this study aims to contribute for the systematized process development when using a technical methodology of data mining in a transportation database analysis. The methodology was formulated by using transportation planning concepts, elaboration of diagnosis and intelligent information system that emphasizes the main data mining concepts. Finally, this methodology was validated while using a rural scholar transportation database populated with a research that was executed in 2.277 Brazilian municipals by the Interdisciplinary Transportation Studies Center – CEFTRU/FNDE(2007a, 2007b).

2. DIAGNOSIS AND TRANSPORTATION PLANNING

In this section, a systematized overview about planning models are introduced for the purpose of providing a better comprehension on how the data mining technique could and must act in a planning process and object diagnosis elaboration.

2.1. Traditional Planning

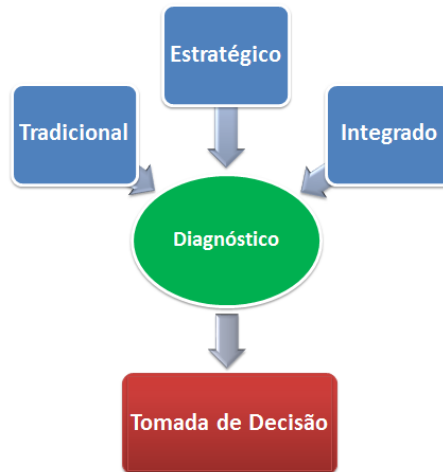
According Papacosta and Provedouros (1993), the continuous traditional planning comes to a systemic vision and is comprised of eight planning steps: (i) goals and objectives definition; (ii) data collection; (iii) analysis of existing conditions; (iv) elaboration of alternatives; (v) alternatives analysis; (vi) evaluation and choice; (vii) implementation; and (viii) continuous valuation. This cyclic process highlights the need of continual valuation.

2.2. Integrated Planning

This model enables the planner to understand the whole planning process and guides him on the build of action plans, implementation, controlling and on efforts valuation in order to transform the planed object. The structure is composed of three hierarchical decision levels: strategic, tactical and operational (Magalhães and Yamashita, 2008).

2.3. Diagnosis

In accordance with Tedesco (2008), the diagnosis will always be part of all planning models, though it is not much discussed in a systematic way as presented in Picture 1. The diagnosis is usually the first step in the planning process without which is not possible to set goals, objectives and define the desired situation. Concisely, it is only possible to identify the problems and find out more accurate solutions for them through a diagnostic that reflects the object status.



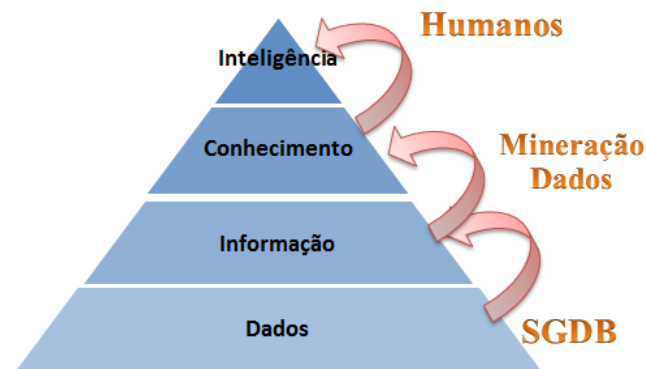
Picture 1: Connection view of the diagnosis. Resource: adaptation of Tedesco (2008)

3. INTELLIGENT INFORMATION SYSTEM

The decision making intelligent systems provides tools and analytical models for the analysis of large amount of data, besides interactive supporting queries for planners to face decision making situations.

3.1. Intelligent Information Systems

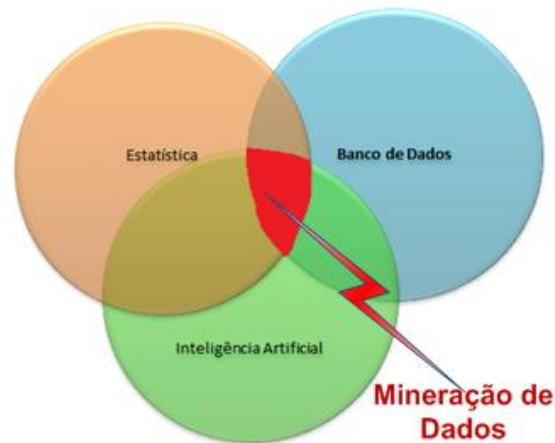
Nonaka and Takeuchi (1997) says that the information can be understood under some perspectives and their classification can involve data, information, knowledge and intelligence under a decision context. These classifications hold different values as shown in Picture 2.



Picture 2: Information pyramid. Resource: Nonaka and Takeuchi adaptation (1997)

3.2. Data mining

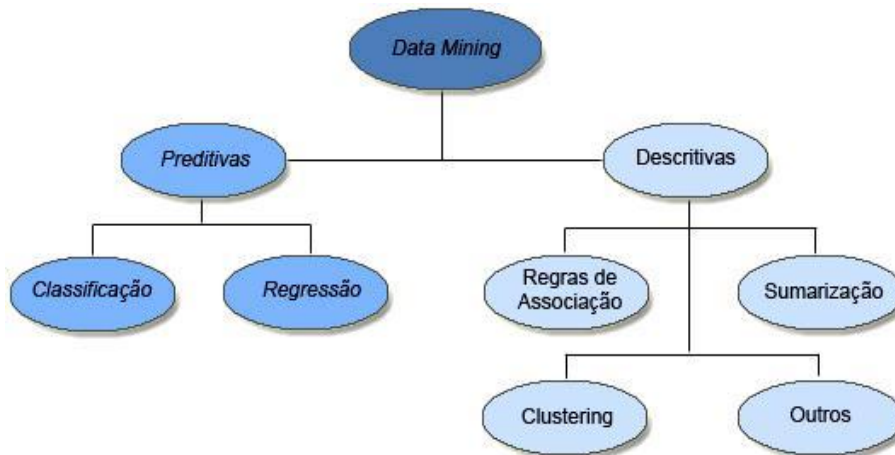
Fayyad (et al., 1996) defines that the data mining process is a single step of a bigger process called knowledge discovery in databases – KDD. In this step, data mining process uses techniques and algorithms from different knowledge areas, mainly artificial intelligence area, database area and statistics area, according to Picture 3.



Picture 3: Multidisciplinary of MD. Resource: Fayyad adaptation *et al* (1996)

3.2.1 Data mining tasks and activities

Data mining systems developed for many kinds of domains are increasing the variety of tasks and activities. Picture 4 introduces the two performed activities vision in the data mining process (Santos and Azevedo, 2005):



Picture 4: Data mining tasks. Resource: Santos and Azevedo adaptation (2005)

- The predictive activities encompass a set of data to detect standards in order to predict interest variables (Santos and Azevedo, 2005).
- The descriptive activities search for interpreted standards that will be detected in the stored data and will give the general characteristics of it (Santos and Azevedo, 2005).

Table 1 presents a resume of the main data mining tasks, according to Dias (2001) explanation.

Tabela 1: Data mining tasks

Task	Description	Example
Classification	Creates a model that can be used in new not classified data to categorize it.	Classify the drivers according to their committed traffic violations.
Estimative (regression)	Estimates the value of a dependent variable as from many independent variables that when grouped produce a result.	Estimate the rural scholar users' transportation travel time.
Association	Determine which factors tend to occur together in the same events.	Determine which traffic violations occur together in a specific area.
Grouping	Divide a set of heterogeneous data distributed in various subsets as more homogeneous among themselves as possible.	Identify the traffic violation driver profile.

Fonte: Dias (2001)

3.2.2 Data mining techniques

Table 2 describes a summary of the most important data mining techniques explained by Dias (2001):

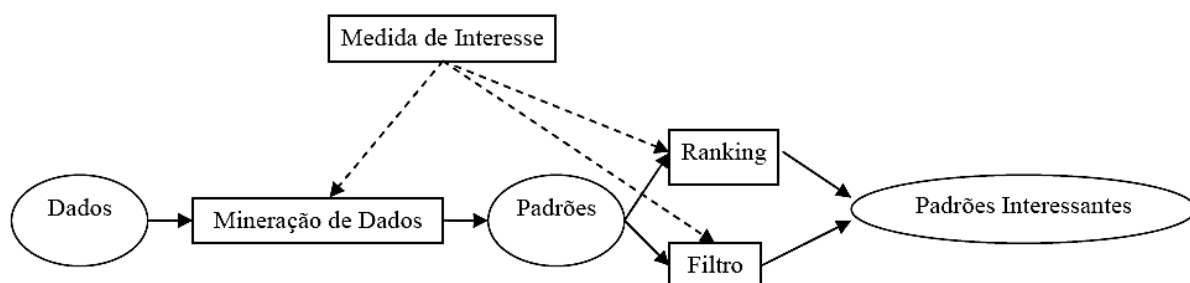
Tabela 2: data mining techniques

Technique	Description	Algorithm
Association rules	Find associations and statistics correlation between data attributes from a set of data.	<i>Apriori, AprioriTid,</i>
Decision tree	Create a tree model that is used to classify new data without testing all attributes.	<i>C5.0, J.48</i>
Generic algorithm	General search and optimization methods, inspired in the evolution process. Find the accurate solution of a specific problem after verifying a enormous number of alternative solutions.	<i>Simple Generic Algorithm and Hillis Algorithm</i>
Artificial neural networks	The model tries to reproduce patterns of the human brain. They are used to solve complex problems that are immerse in a big quantity of collected data.	<i>Perceptron, MLP Network, Hopfield Network, BAM Network</i>

Resource: Dias (2001)

3.2.3 Comprehension of association rules measures of interest

For Geng and Hamilton (2006), the interest objective measures can be used in three different ways inside the data mining process as Picture 5 illustrates.



Picture 1: Interest objective measures. Resource: Geng and Hamilton (2006)

Some objective measures of interest that will be used in this study will be shown in the following.

Support: corresponds the frequency that A and B happens in the entire database (Witten and Frank, 1999), as equation 1.

$$\text{Suporte } (A \Rightarrow B) = \frac{\text{Sup}(A \cup B)}{N} \quad (1)$$

wherein: Suporte $(A \Rightarrow B)$: objective measures of interest real support;
Sup $(A \cup B)$: number of records containing A and B;
N: total number of records in the database.

Reliability: corresponds to the frequency that B occurs among the instances that it is contained (Witten and Frank, 1999), as equation 2.

$$\text{Confiança } (A \Rightarrow B) = \frac{\text{Sup}(A \cup B)}{\text{Sup}(A)} \quad (2)$$

wherein: Confiança $(A \Rightarrow B)$: reliable objective measures of interest;
Sup $(A \cup B)$: number of records containing A and B;
Sup (A) : number of records that contains A;

Suporte Esperado: O suporte esperado é computado, baseado no suporte dos itens que compõem a regra (Brin *et al.*, 1997), conforme a equação 3.

$$\text{SupEsp } (A \Rightarrow B) = \text{Sup}(A) * \text{Sup}(B) \quad (3)$$

wherein: SupEsp $(A \Rightarrow B)$: expected objective measure of interest;
Sup (A) : number of records that contains A;
Sup (B) : number of records that contains B;

Lift: This measure is one of the most used measures in the valuation of two items dependency (Brin *et al.*, 1997), as equation 4.

$$\text{Lift } (A \Rightarrow B) = \frac{\text{Conf}(A \cup B)}{\text{Sup}(B)} = \frac{\text{Sup}(A \cup B)}{\text{Sup}(A) * \text{Sup}(B)} \quad (4)$$

wherein: Lift $(A \Rightarrow B)$: lift objective measure of interest;
Sup $(A \cup B)$: number of records containing A and B;
Sup (A) : number of records that contains A;
Sup (B) : number of records that contains B;

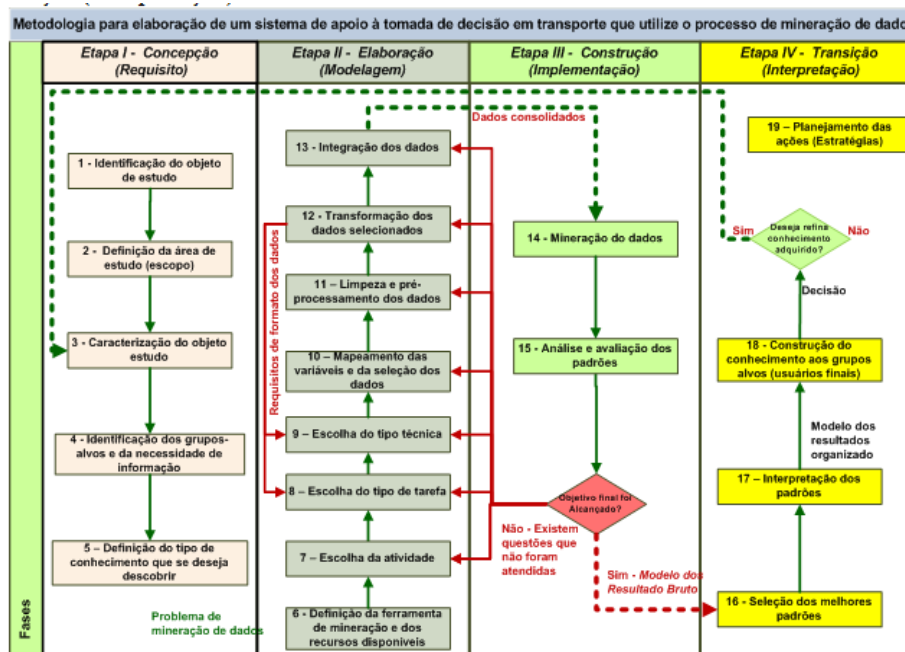
3.3. Tools

The software used in this study was the development of the complete WEKA library and new requirements were implemented in the *a priori* association rule algorithm. Its graphic interface is easier to use because the information about the file selector option and the algorithms are in each screen for a specific task.

Corresponding what was exposed above, it's clear the importance of developer a methodology using the data mining process in the transportation database analysis, for the propose of extract patterns and significant rules to support managers during transportation planning process to decision making.

4. TRANSPORTATION DECISION MAKING METHODOLOGY USING THE DATA MINING PROCESS

The methodology proposed in Picture 6 is composed of 19 activities that are separated into 4 large steps: inception, elaboration, construction and transition.

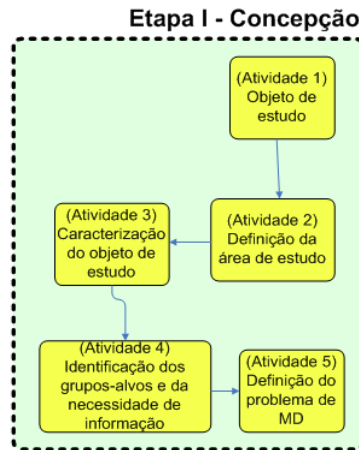


Picture 2: Proposed methodology

Picture 6 represents the functional methodology view when describing the logic processes. Each process is composed of a sequence of tasks and decisions that control when and how the tasks will be executed. This logic representation is more familiarized as flowchart, because it helps to understand the tasks sequence that is embedded in the four steps, which are detailed in the following topics.

4.1. Etapa I – Conceção (requisitos)

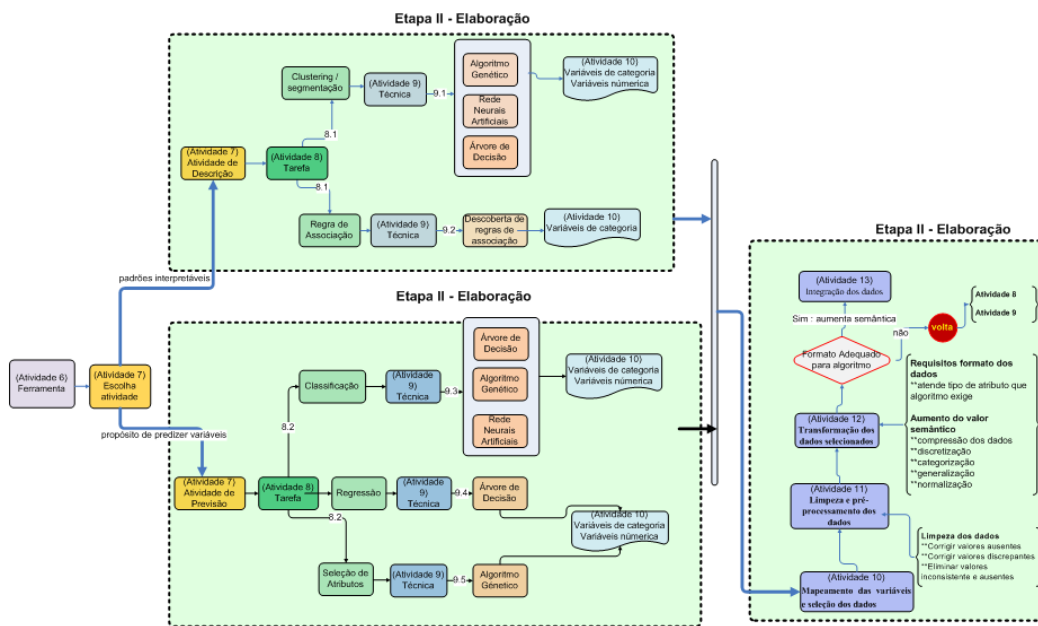
This step is the most important and determinant moment. It's formed by five activities: clear definition of the study object, data mining problem uptake, study area delimitation, identification of target groups and their information needs more accurate for each profile inside the decision making process, as Picture 7 shows.



Picture 7: Inception methodology step

4.2. Second Step: Elaboration (modeling)

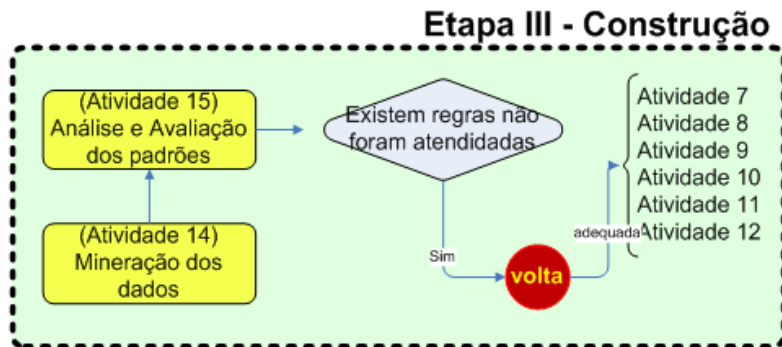
The most expensive part of the process, this step is formed by seven activities that consist in choosing a tool for the type of activity that will be followed in the data mining process; in cleaning and preprocessing the selected data; in data transforming and data integrating; and in generating a selected data file readable by the tool format, as seeing in Picture 8.



Picture 8: Methodology's Elaboration Step

4.3. Third Step - Construction (implementation)

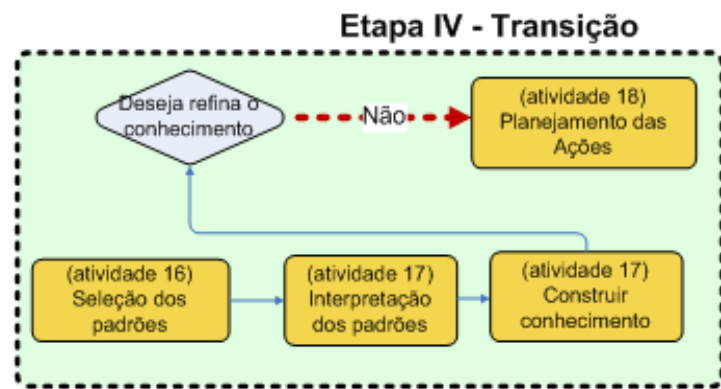
Construction step is composed by two activities. The first on is the method application to data patterns extraction with the view to find the better algorithm parameters appliance for the specified job. After that, the post-processing activity is performed where the goal is identify more interesting patterns according to business domain, as Figure 9 shows.



Picture 9: Methodology's Construction Step

4.4. Fifth Step – Transition (interpretation)

In Transition step, the patters generated by a determined business context are interpreted and evaluated. It is structured by five activities to generate information, useful and understandable knowledge to planners and managers subsidy in the action plan elaboration, as presented in Picture 10.



Picture 10: Methodology's Transition Step

5. STUDY CASE: RURAL SCHOLAR TRANSPORTATION CHARACTERIZATION USING THE DATA MINING PROCESS

Ensuring education access to students that lives in rural areas is a fundamental government rule for social inclusion. Herewith, education access is a constitution guarantee, as presented in article 206, numeral I and in article 208, numeral VII of the Brazilian Federal Constitution of 1998. The articles establish that the State must: assures by supplemental government programs of educational materials, transportation, food and health assistance

(BRASIL, 1998), education access and provide a way for students to reach at the school.

In accordance with Carvalho (2011), in last few years the provision of rural scholar transportation quality are being observed by the Brazilian Public Power. Meanwhile, over the years, a essential service that directly affects education guarantee and students school performance as rural scholar transportation service received little government investment as seen through precarious access of education units.

5.1. Scholar Transportation System Concept

Joining definitions about the concept of the transportation system formulated by some referential transportation authors, Magalhães (2010) realized that the transportation system is a way to reach an end.

5.2. Proposed Methodology Application

In order to simplify the application of the study case, a restriction was implemented using the data mining technique, applied to analyze the data collected in a web research of rural scholar transportation conducted by the CEFTRU/FNDE (2007b, 2007b) in 5.564 Brazilian municipals. From all Brazilian municipals that are distributed across Brazilian States, only 2.277 answer all formulary items of the research questioner on time.

5.2.1 First Step – Inception (requirements)

The rural scholar transportation – RST is considered as the study object. It's provided by the public power for the student exclusive use which can be executed by the prefecture or by as an outsourced service. The rural scholar transportation as a set of interrelated elements common aims to move the rural resident students or rural students from their homes to a educational establishment (CEFTRU/FNDE, 2008a, 2008b).

This study's goal was to characterize the rural scholar transportation in a united and complementary way by means of three objectives: service, customer and resources (CEFTRU/FNDE, 2007a, 2007b). To TER's management, against each municipals specific reality, understand practices and procedures to be adopted for the municipals was considered as the data mining problem.

5.2.2 Second Step – Elaboration (modeling)

The descriptive activity choice to identify correlation in the dataset was done in accordance with the goals of step I of finding associations between adopted practices and procedures for rural scholar transportation management. On the other hand, the association task was chosen to reach the proposed objective. To the authors Witten and Frank (1999), the

association task objective is to establish concept rules, in other words, situations that tend to occur in the same transaction. Then, the presence of some concepts in a transaction implies in the presence of others concepts in the same transaction, identifying, thereby a relation or a tendency.

Corresponding to the described methodology, the association rules was the chosen technique to be used with the apriori algorithm between all options described in this study. Table 3 shows fourteen relevant variables used in selection phase, extraction phase and transformation phase of selected data.

Table 3: used to data selection, extraction and transformation

Information	Element	Description
1 - Scholar Quantity	Characterization - customer	Quantity of schools attended by the transportation service.
2 - Students Quantity	Characterization - customer	Quantity of students attended by the transportation service.
3 - Vehicles Quantity	Characterization - service	Quantity of vehicles used for the transportation service.
4 - Exclusive Vehicles	Characterization - service	Scholar transportation with exclusive vehicles provided by the municipal – student or passengers.
5 - Vehicle Property	Characterization - service	The vehicle property used in the scholar transportation – own property or outsourced property.
6 - Founding Resource	Founding - resource	Resources founding used to finance the scholar transportation provided by the municipal.
7 - Payment Criteria	Destination - resource	Payment criteria used for outsourced scholar transportation.
8 - Elementary and middle schools' students	Identification	Students attended by the scholar transportation that have typical development.
9 – RST for Special Necessities rural students	Identification	Existence of students with special educational necessities that uses the transportation service.
10 – Vehicles used by another finality	Characterization - service	Exclusive scholar transportation vehicles are being used for another end.
11 – Service offered during scholar calendar	Characterization - service	Scholar transportation is offered during all scholar calendar.
12 – RST Adapted Vehicle	Characterization - service	Vehicle is adapted for scholar transportation.
13 – RST accompanied routine exists	Characterization - service	Any kind of routine accompaniment for scholar transportation service exists.
14 – Municipal regulation exists	Characterization - regulation	Municipal regulation for scholar transportation existence.

During data extraction data cleaning using basic operations of missing information deletion intended to be necessary. The apriori algorithm only accepts nominal fields, it means that the algorithm do not work with quantity values. This way, transform the data into another category using the discretization technique applied in some fields was necessary. Only the

filters were used in Structured Query Language – SQL instructions.

Finally, as the chosen tool works with an owned format, a data base was generated taking as the premise three fundamental elements for the rural scholar transportation characterization: service, customer and resources as shown in Table 3. An overall representation of TER's characteristic elements was possible to be expressed. Then, the ARFF file extension describes the attribute domain. It is an ASCII file to define attributes and their values: 'attribute=value'.

5.2.3 Step III – Construction (implementation)

As described on step II, the apriori algorithm with some logical structure modifications to generate the association rule was applied. This alteration was intended to implement the specification of objective measure of interest: support, reliability and lift. This implementation is necessary to avoid data mining excessive number of obvious association rules in the dataset.

5.2.4 Step IV – Transition (results interpretation)

As result of the selected dataset in step II, 45 association rules were presented by the program. Those association rules were transported to an excel spreadsheet file to simplify the analysis and to discover the more significant association rules according to each measure of interest. The generated association rules have the values of 'real calculated support' bigger or equals the 'minimum support' of 30% and less or equals the 'maximum support'. The 'calculated reliability' value is bigger or equals the 'minimum reliability' of 30% and less or equals the 'maximum reliability' of 100%. And, the 'calculated lift' value is bigger or equals the 'minimum lift' of 1,10. The lift use as a positive dependency is the cause of little number of rules generated.

Two ways of data interpretation as (i) Analytical analysis to perform extracted patters interpretation to transform data into information and useful knowledge; (ii) descriptive statistics analysis to describe the study object producing schematized results formatted in graphics and tables.

Table 4 describes the structure of association rules number 1 as following:

- Rule body
 - *'exclusive vehicle for students'*;
 - *'vehicle is adapted'*;
 - *'existing regulation'*.
- Result
 - *'quantity of vehicles used'* ;
 - *'elementary and middle school's students'*.

Table 4: Structure of association rule number 1

R: 1	'exclusive vehicles for students = yes', 'is vehicle adapted = no', 'is there any regulation = no' <table border="1"> <tr> <td>Measure of interest</td> </tr> <tr> <td>'support (R)' = 41%</td> </tr> <tr> <td>'reliability(R)' = 61%</td> </tr> <tr> <td>'lift(R)' = 1,1</td> </tr> </table>	Measure of interest	'support (R)' = 41%	'reliability(R)' = 61%	'lift(R)' = 1,1	⇒	'quantity of vehicles used < 5' 'student is in elementary or middle school=yes'
Measure of interest							
'support (R)' = 41%							
'reliability(R)' = 61%							
'lift(R)' = 1,1							

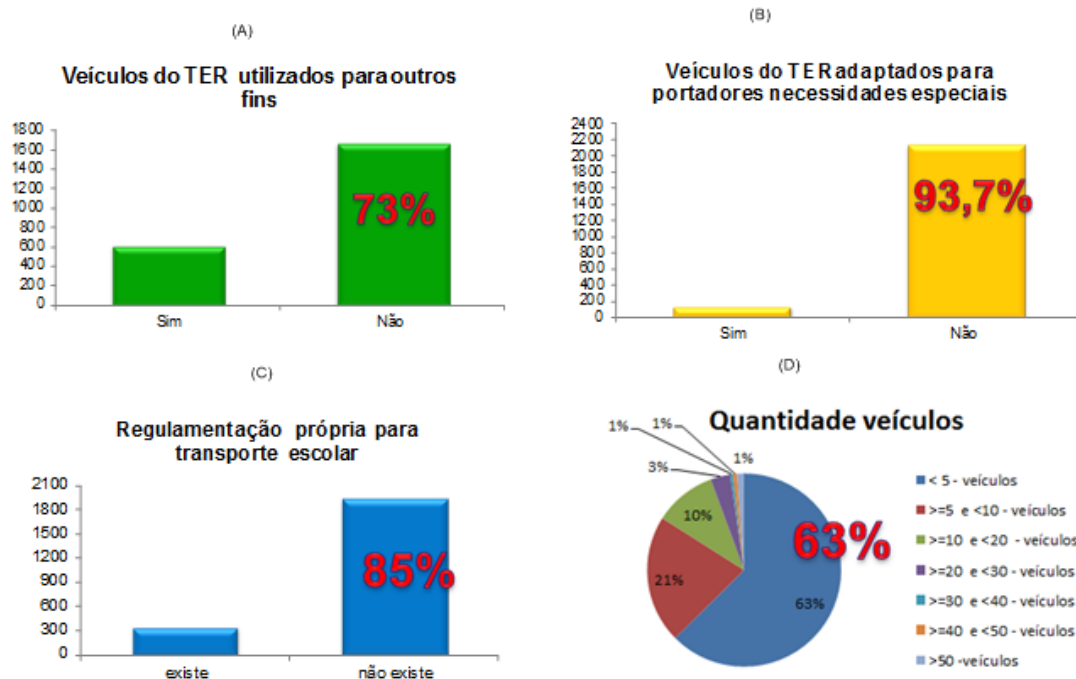
(i) Analytical analysis – Data mining: association rule number 1

- The *support measure* value indicates that 41% of all municipals that do not have any owned regulation use less than five non adapted vehicles for special needs passengers of the exclusive rural scholar transportation for students coursing elementary or middle school.
- The *reliability measure* indicates that the probability of one municipal that have any owned regulation use less than five non adapted vehicles for special needs passengers for exclusive rural scholar transportation of the students coursing elementary or middle school is 61%.
- The *lift measure* indicates that the municipals that uses less than five vehicles for rural scholar transportation of students coursing elementary or middle school is 1,1 times bigger than the municipals that uses non adapted vehicles for special needs passengers; that uses exclusive vehicles and that do not have any owed regulation.

This association rule identified facts that tend to occur together in a single transaction. Then, the presence of them in a transaction implied in the presence of others in the same transaction identifying then a tendency or a relationship between them. This rule presents correlation between the variables, in other words, if a manager offers a transportation service to a customer that have students with special needs, it means that the municipal had a non proper action.

(ii) Análise Descriptive analysis – descriptive statistics: variables of rule number 1

Picture 11-A shows about 27% of the municipals that uses vehicles for others ends when those vehicles are not being used for rural scholar transportation. Picture 11-B show about 93% of the municipals that uses non adapted vehicles for special needs passengers. Picture 11-C shows about 85% of the municipals that declared not having any own regulation for scholar transportation. Picture 11-D show about 63% of the municipals that uses less than five vehicles for rural scholar transportation. (CEFTRU/FNDE, 2007b, 2007c).



Picture 3: Gráficos Graphics of rule number 1 variables. Resource: CEFTRU/FNDE (2007b, 2007c)

6. CONCLUSION

This study had as a study object to propose a methodology that uses the data mining process to explore and analyze databases. The goal was to discover patterns and significant and adequate rules for information needs for decision making activity support during all planning, management and control transportation processes.

The analysis results of a database, commonly the descriptive statistics analysis is used to organize data in tables and graphics. This study used another analysis, the analytical analysis that seek to transform data into information, in other words, transform data in comprehensible and useful information to give supporting for decision making activity on planning, on managing and on controlling. Finally, in a complete analysis environment is necessary to get results from the two kinds of analysis. It means that the analyses are complementary and not overlapped.

From the results presented in this study, a recommendation to continue this studying research for development of new data mining methodologies with graphics representation and spatial database support (for example, application of geographical elements, images of remote sensing for transportation system analysis). It's important to describe some limitations of this study as: (i) little scientific studies about data mining methodologies and (ii) use of data from secondary resource declared by the managers of rural scholar transportation.

7. REFERÊNCIAS BIBLIOGRÁFICAS

- BRASIL (1988) Constituição da República Federativa do Brasil. Brasília, DF: Senado. Diário Oficial da República Federativa do Brasil, Brasília, 5 de outubro de 1988.
- BRIN, S. et. al.(1997). Dynamic Itemset Counting and Implication Rules for Market Basket Data. Estados Unidos: ACM SIGMOD.
- CARVALHO, Willer Luciano. (2011) Metodologia de Análise para a localização de escola em áreas rurais, Tese de Doutorado. Faculdade de Tecnologia, Departamento de Engenharia Civil e Ambiental, Universidade de Brasília, DF, 215p.
- CEFTRU/FNDE (2007a) Centro Interdisciplinar de Estudos em Transporte e Fundo Nacional de desenvolvimento da Educação. Projeto: Transporte Escolar Rural - volume I - Metodologia de caracterização do transporte escolar rural. Brasília, DF. 2007.
- CEFTRU/FNDE (2007b) Centro Interdisciplinar de Estudos em Transporte e Fundo Nacional de desenvolvimento da Educação. Projeto: Transporte Escolar Rural - volume II - Levantamento de dados para a caracterização do transporte escolar – questionário web. Brasília, DF. 2007.
- CEFTRU/FNDE (2007c) Centro Interdisciplinar de Estudos em Transporte e Fundo Nacional de desenvolvimento da Educação. Projeto: Transporte Escolar Rural - volume III – Tombo I - Caracterização do transporte escolar nos municípios visitados. Brasília, DF. 2007.
- CEFTRU/FNDE (2008a) Centro Interdisciplinar de Estudos em Transporte e Fundo Nacional de desenvolvimento da Educação. Projeto: Transporte Escolar Rural – volume I - Diagnóstico do Transporte Escolar Rural. Brasília, DF. 2008.
- CEFTRU/FNDE (2008b) Centro Interdisciplinar de Estudos em Transporte e Fundo Nacional de desenvolvimento da Educação. Projeto: Transporte Escolar Rural – volume II - Diagnóstico do Transporte Escolar Rural. Brasília, DF. 2008.
- DIAS, Maria Madalena. (2001) Um modelo de formalização do processo de desenvolvimento de sistemas de descoberta de conhecimento em banco de dados. Tese Doutorado - Curso de Pós-Graduação em Engenharia de produção, Universidade Federal de Santa Catarina.
- FAYYAD, U. M.; Piatetsky-Shapiro, G. e Smyth, P. (1996) From Data Mining to Knowledge Discovery: An Overview. In: Advances in Knowledge Discovery and Data Mining, AAAI Press.
- GENG, L. e HAMILTON, H. J.(2006) Interestingness Measures for Data Mining: A Survey. ACM Computing Surveys, v. 38, n. 3.
- MAGALHÃES, M. T. Q. (2010) Fundamentos para a pesquisa em transporte: reflexões filosóficas e contribuições da ontologia de Mário Bunge. Tese (Doutorado em Transportes) - Faculdade de Tecnologia, Universidade de Brasília, Brasília.
- MAGALHÃES, M. T. Q. e YAMASHITA, Y. (2008) Repensando o Planejamento. Texto para discussão n.4. CEFTRU/ UnB: Brasília, 2008.
- NONAKA, I e TAKEUCHI, H. (1997) Criação de conhecimento na empresa. Rio de Janeiro: Campus.
- PAPACOSTAS C. S. e PROVEDOUROS, P. D. (1993) Transportation Engineering and Planning. 2ª. ed. New Jersey: Prentice Hall.

- SANTOS, M. Filipe e AZEVEDO S. Carla. (2005) Data Mining - Descoberta de Conhecimento em Bases de Dados. Portugal : Editora FCA.
- TEDESCO, G. M. I. (2008) Metodologia para Elaboração do Diagnóstico de um Sistema de Transporte. Brasília. 2008. 215p. Dissertação (Mestrado em Engenharia de Transporte) — Faculdade de Tecnologia, Universidade de Brasília.
- WEKA - University of Waikato. Weka 3.6 – Machine Learning Software in Java. Disponível em < <http://www.cs.waikato.ac.nz/ml/weka> > Access: 10/02/2011.
- WITTEN, I. H., AND FRANK E. (1999) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco, 1999.