ISSN 1648-4142 TRANSPORT
http:/www.vtu.lt/english/editions

TRANSPORT – 2003, Vol XVIII, No 5, 216–223

DATA MINING APPLICATIONS IN TRANSPORTATION ENGINEERING

Sudhir Kumar Barai

Dept of Civil Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721 302 India
E-mail: skbarai@civil.iitkgp.ernet.in
Phone: 91-3222-283408; FAX: 91-3222-282254

Received 2003 05 28; accepted 2003 09 01

Abstract. Data mining is the extraction of implicit, previously unknown and potentially useful information from data. In recent time, data mining studies have been carried out in many engineering disciplines. In this paper the background of data mining and tools is introduced. Further applications of data mining to transportation engineering problems are reviewed. The application of data mining for typical example of 'Vehicle Crash Study' is demonstrated using commercially available data mining tool. The paper highlights the potential of data mining tool application in transportation engineering sector.

Keywords: Data Mining, GIS, GPS, Pavement Management, Road accidents, Traffic Management, Transportation.

1. Introduction

Data mining is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in dataset. This process helps in extracting and refining useful knowledge from large datasets. The extracted information can be used to form a prediction or classification model, identify trends and associations, refine an existing model, or provide a summary of the datasets being mined. Numerous data mining techniques of various types such as rule induction, neural networks, and conceptual clustering have been developed and used individually in domains ranging from space data analysis to financial analysis [1]. A recent review by Kohavi [2] states that data mining serves two goals namely *Insight* and *Prediction*. Insight leads to identifying patterns and trends that are useful. Prediction leads to identifying a model that gives reliable prediction based on input data.

Obviously, the nature of data is critical to the success of data mining application. The nature of the data is related to its *source, utility, behavior* and *description*. Source of data can be online or off-line from static or dynamic systems. Data utility can be for analysis, design or diagnosis. Behavior of data can be discrete or continuous. Data description can be in a quantitative or qualitative form. A quantitative nature of the data depends on the number of data points available for an application. A qualitative nature of the data demands answers to many questions such as, Are they sparse or dense? Are they in a raw or clean form? Are they representative of the application domain? Are they noisy? Do they contain missing values? Researchers working in the field of scientific data mining have addressed an issue of *insight* such as novelty detection, anomalies and faults in experimental data for classification and regression problems [3]. They are commonly addressed for pattern recognition, image analysis, process monitoring and control, and fault diagnostics and so on.

In the field of transportation engineering large amounts of data are generated during studies on traffic management, accidents analysis, pavement conditions, roadway feature inventory, traffic signals and signal inventory, bridge maintenance, road characteristics inventory etc. Based on these data, decision-makers arrive at decision to solve a respective problem. Decision-makers are always on look out for ways to ease the pain in obtaining access to and applying disparate datasets. The basic requirements include the ability to identify what data is available, determine the characteristics of the data, extract the data of interest, transform the data into formats necessary for the application. In real life the situation of transportation domain, diverse fields of data need to be collected to integrate and to arrive at the solutions. Recent research study in the field of data mining approach has opened a new horizon for decision-makers of transportation engineers [4].

The main objectives of the paper are as follows:

- To introduce briefly the data mining and commercially available software
- To review data mining application in the context of transportation engineering

- To demonstrate a data mining problem with commercially available software for 'vehicle crash data' problem.

2. Background On Data Mining And Knowledge Discovery

There is a broad spectrum of engineering problems where computational intelligence is becoming an essential part in many advanced systems. Such problems arise in data processing, which is faced with huge data explosion, due to automatic data collection systems and the possibility for combining data from many sources over data networks. Hence new techniques for extracting important knowledge from the raw data are required to efficiently handle such data. Data mining is a step in the *Knowledge Discovery and Data Mining* process consisting of particular data mining algorithms that, under some acceptable computational efficiency limitations, produces a particular enumeration of patterns [1].

Basic steps involved in *data mining and knowledge discover*y (refer Fig 1) are as follows and detail explanation can be referred elsewhere [1].

1. Understanding of the application domain
2. Collection of target dataset
3. Data cleaning and pre-processing
4. Data Warehousing
5. Selection of task relevant data selection
6. Selection of data mining task
7. Selection of data mining tool - Artificial neural networks, Genetic Algorithms, Decision trees, Nearest neighbor method, Rule induction, Data visualization
8. Data mining - relationship identification - Classes, Clusters, Associations, Sequential patterns
9. Interpretation of results
10. Consolidation of discovered knowledge

3. Commercially Available Data Mining Tools

Recently, King et al. [5], Elder IV and Abbott [6] and Abbott et al. [7] carried out systematic studies on the performance of some of the popular data mining tools. Based on the results of their studies, the strong and weak points of these tools have been summarized in Table 1. General observations from the studies were based on the following various aspects.

- Platforms Supported
- Algorithms Included - Decision trees, Neural networks, Other
- Data Input and Model Output Options
- Usability Ratings
- Visualization Capabilities
- Modeling Automation Methods

The comprehensive study had the following observations for data miners.

- Data mining tools can: enhance inference process, speed up design cycle
- Data mining tools can not: substitute for statistical and domain expertise
- Users are advised to: get training on tools, be alert for product upgrades

For data miners more details about the data mining tools can be found elsewhere [8].

4. Potential Problems Of Data Mining In Transportation Engineering: Brief Review

Traffic Management

For many years, many researchers have been developing a unique approach to road traffic management and congestion control. The system could be built to classify vehicles into different categories. Road sensor will collect data and such dataset would me mined to classify

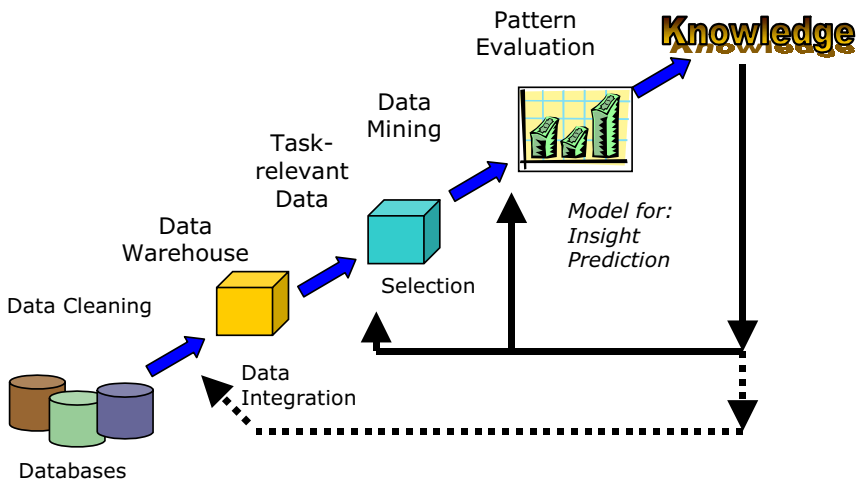

Fig 1. Data mining and knowledge discovery process

Table 1. Commercially Available Data Mining Software [5–7]

Product	Company	Strengths	Weaknesses
Clementine	Integral Sol., Ltd.	visual interface; algorithm breadth	scalability
Darwin	Thinking Machines, Corp.	efficient client-server; intuitive interface options	no unsupervised; limited visualization
DataCruncher	DataMind	ease of use	single algorithm
Enterprise Miner	SAS Institute	depth of algorithms; visual interface	harder to use; new product issues
GainSmarts	Urban Science	data trans formations , built on SAS; algorithm option depth	no unsupervised; limited visualization
Intelligent Miner	IBM	algorithm breadth; graphical tree/cluster output	few algorithm options ; no automation
MineSet	Silicon Graphics, Inc.	data visualization	few algorithms; no model export
Model 1	Group 1/Unica Technologies	ease of use; automated model discovery	really a vertical tool
ModelQuest	AbTech Corp.	breadth of algorithms	some non-intuitive interface options
PRW	Unica Technologies, Inc.	extensive algorithms; automated model selection	limited visualization
CART	Salford Systems	depth of tree options	difficult file I/O; limited visualization
NeuroShell	Ward Systems Group, Inc.	multiple neural network architectures	unorthodox interface; only neural networks
OLPARS	PAR Government Systems	multiple statistical algorithms; class-based visualization	dated interface; difficult file I/O
Scenario	Cognos	ease of use	narrow analysis path
See5	RuleQuest Research	depth of tree options	limited visualization; few data options
S-Plus	MathSoft	depth of algorithms; visualization; programmable/extendable	limited inductive methods; steep learning curve
WizWhy	WizSoft	ease of use ; ease of model understanding	limited visualization

the vehicles. Further, there is a need on data mining research on the fundamental relationship of traffic flows and their interpretation under congested road conditions. Such application of data mining will reduce the cost of transportation planning significantly [9]. The traffic management problems using data mining approach may demand answer for the following various issues, such as new data mining approaches to automate the process of knowledge acquisition and representation, an appropriate structure of a data warehouse to support traffic management, and new approaches for improved traffic data presentations and multimedia based visualization techniques.

Monitoring drowsy drivers

Data regarding general characteristic of a driver's behavior in the sleep could be captured using array of sensors in a driver's cabin. Then, the data mining techniques could be applied to analyze such data, to determine when truck drivers are likely to fall asleep and alert them by an alarm to avoid accidents [9].

Road Accidents Analysis

While designing the road networks, data related to dangerous and safe stretches are collected. This helps in planning road improvement schemes. Accident records can be misleading as the frequency of accidents varies considerably. The problem of getting reliable estimates

of the long-term road accident frequencies at individual road location is a challenging problem. The data mining of previously collected data of road networks will help in identifying high risk sites inspite of fluctuating frequency of accidents. The data mining exercise could be carried out using the cause and effect process data related to accident frequencies, environmental and other variables [10].

In other studies [11], varieties of data mining techniques have been applied by the participating partners (e.g. Leuven, Prague, Ljubljana, Bristol). Highlights of the data mining studies on road accident analysis were as follows:

- An innovative visualization method was developed to animate the development of accidents by location over time. The method helped in locating data quality issues regarding grid references.
- Association rules approach was applied to find associations between road numbers and particular classes of accidents
- Text mining technology and subgroup discovery approaches helped in determining common kinds of accidents.
- Dynamic subgroup discovery approach highlighted certain data quality issues.

Pavement Management Data

Conventional Pavement Management System (PMS) generates a large amount of data. The analysis of such datasets has been great challenge to many highway agencies to take major decision for pavement maintenance and rehabilitation. In recent time, data mining technique has been used to pavement serviceability ratings (PSR), as an indirect way of obtaining the remaining life of pavements. Recently, data mining study was carried out for the parameters such as were the present serviceability rating, rutting, ride quality, condition, cracking, year the pavement was last worked, AADT, overlays, original surface, and surface type [4]. Further, study can be carried out considering parameters like the percentage of trucks and passenger vehicles, design life of each pavement, type of pavement (i.e. concrete/asphalt, and if concrete, reinforced/un-reinforced or continuously reinforced), and environment. Great potential lies ahead for data miners for PMS data.

Geographic Information Systems for Transportation Data

In Geographic information systems (GIS) for transportation, interconnected hardware, software, data, people, organizations and institutional arrangements (for collecting, storing, analyzing and communicating particular types of information about the earth) plays a crucial role in generating huge amount of data [12]. In application of GIS for transportation network, complex data (e.g. logical, physical, real and virtual world) as given below (Fig 2) are generated.

The above-discussed GIS based multifaceted transportation data has complex relationship with each other. Data mining can be an excellent tool to identify these complex relationships between data nature of logical, physical, real and virtual world

Global Position Systems Data

An automated technique such as Global Position System (GPS) has been advocated for navigation applications in vehicles, and generating detailed maps against the manual lane measurements. GPS generates position traces with differential corrections. The size of such data is too large and obtaining a refined map of these traces has been a challenging task. Data mining approach has been proposed to generate such refined map from GPS data. This approach helps in lanekeeping and convenience applications such as lane-changing advice [13].

Roadway Videologs Spatial Data

Video-logging systems use data collection vehicles to collect the data on pavements and roadside structures and to take videos of the right-of-way. The video information used by highway agencies is stored in an analog format and located at specific locations. The storage media include tapes, films, and laser disks. Engineering site data are stored in separate databases. The objectives of such data collection are to obtain quickly and cost-effectively acquiring spatially referenced geometric and inventories data for roads and highways and to demonstrate the use of those data for practical planning application. With the possibility of accurate spatial data representation in GIS, there is need of extracting the necessary information from videologs. Here data mining has potential utility to extract the important features from videologs [14].

	Logical	**Physical**
Real World	***Legal definitions*** •Route •State trunk network •County trunk network •Street network •Political boundary	***Actual facilities*** •Highways •Roads •Interchanges •Intersections
Virtual World	***Data Structures*** •Networks •Chains •Links •Nodes •Lattices	***Data Values*** •Lines •Points •Polylines •Polygons •Attributes

Fig 2. Data characteristics in GIS applications of transportation engineering [12]

Table 2. Typical Dataset of Crash Record

make	Model	Year	Doors	Wt	Size	Protection	D/P	Head IC	Chest decel	L Leg	R Leg
Acura	Integra	87	2	2350	lt	manual belts	Driver	599	35	791	262
Audi	80	89	4	2790	comp	manual belts	Driver	600	49	168	1871
BMW	325i	90	2	2862	comp	d airbag	Driver	1036	56	865	
Buick	Elect. Park Ave	88	4	3360	med	manual belts	Driver	1467	54	712	1366
Buick	Regal	88	2	3210	med	passive belts	Driver	880	50	996	642
Cadillac	De Ville	90	4	3500	hev	d airbag	Driver	423	39	541	1629
Chevrolet	Astro	88		3787	van	manual belts	Driver	1603	72	1572	700
Chrysler	Fifth Ave	89	4	3820	hev	d airbag	Driver	786	43	1132	667
Daihatsu	Charade	88	2	1820	mini	manual belts	Driver	768	43	574	598
Dodge	colt	88	4	2348	lt	manual belts	Driver	1354	53	844	461
Eagle	Medallion	89	4	2740	comp	Motorized belts	Driver	745	41	1721	1574
Ford	Aerostar	87		3013	van	manual belts	Driver	1568	49	286	590
Geo	Metro	89	2	1590	mini	manual belts	Driver	951		789	458
Honda	Accord	87	2	2440	lt	passive belts	Driver	769	46	863	1244
Hyundai	Excel	87	4	2200	lt	passive belts	Driver	757	54	2408	1187
Infiniti	M-30	90	2	3370	med	d airbag	Driver	466	43	936	844
Isuzu	Amigo	90	2	2900	mpv	manual belts	Driver	996	56	388	501
Jeep	Cherokee 4x4	89	4	3284	mpv	manual belts	Driver	968	69	401	540
Lexus	ES250	90	4	3280	med	d airbag	Driver	992	55	1102	1012

Road Roughness Data Analysis

Measurement of road roughness helps in getting an over all idea about the quality of a pavement, road user satisfaction and vehicles operating costs. In this regard, highway engineers carry out systematic roughness surveys with the help of roughness meters. During survey a sizeable amount of data is collected for parameters like Road ID, Local referencing points, road roughness etc. [15]. The data mining tool can be used in determining the relationship between the location, road user satisfaction and road roughness.

In this section, various potential applications of data mining have been reviewed in the context of transportation engineering problems. Next section demonstrates the data mining application for crash test data.

5. Example - Crash Test Database

For the demonstration of data mining application in transportation engineering field, we have collected dataset on ***Crash Test Dummies***. Automobiles with dummies in the driver and front passenger seats were crashed into a wall at 35 miles per hour. National Transportation Safety Administration collected the information how the crash affected the dummies. The injury variables describe the *extent of head injuries*, *chest deceleration*, and *left and*

right femur load. The dataset also contains information on the type and safety features of each crashed car [16]. The dataset had 352 records and following variables. Typical dataset is given in Table 2.

- Make: Car make
- Model: Model of that car
- Year: Year of the car
- Doors: Number of doors in the car
- Wt: Weight in pounds
- Size: A categorical variable to classify the cars to a type (light, minivan)
- Protection: Kind of protection (seat belt, air bag, etc.)
- D/P: Whether the dummy is in the Driver or Passenger seat
- Head_IC: Head injury criterion
- Chest_decel: Chest deceleration
- L_Leg: Left femur load
- R_Leg: Right femur load

Using these records, first eight variables as dependent variables, knowledge rules were generated for Head Injury Criterion, Chest deceleration, Left femur load and Right femur load using commercially available data mining software - WizRule 4.01 and WizWhy 4.02. The details and features about this software are available at [17].

Typical screen shot of WizRule is shown in Fig 3.

WizRule/WizWhy reads the data and allows fine-tuning the analysis parameters such as "*minimum probability of if-then rules*" and "*minimum number of cases of a rule.*" User has a control over defining exactly which types of rules *WizRule/WizWhy* should look for.

WizRule/WizWhy identifies the rules based on the data and discovers the deviating rules. The deviating rules with the highest degree of unlikelihood are listed as *suspected errors. If-then* rules generated using WizWhy are as follows in Table 3.

The "*Probability*" in if-then rules is defined as the ratio between the number of records in which the condition(s) and the result hold, and the corresponding number of records in which the condition(s) hold with or without the result. The "*Significance Level*" means the degree of the rule's validity. It is equal to 1 minus the "*error probability*", which quantifies the probability that the rule exists accidentally in the data under analysis.

Knowledge rules generated from a large size of dataset using data mining tools can become an integral part of Decision Support System or Knowledge Based Expert Systems.

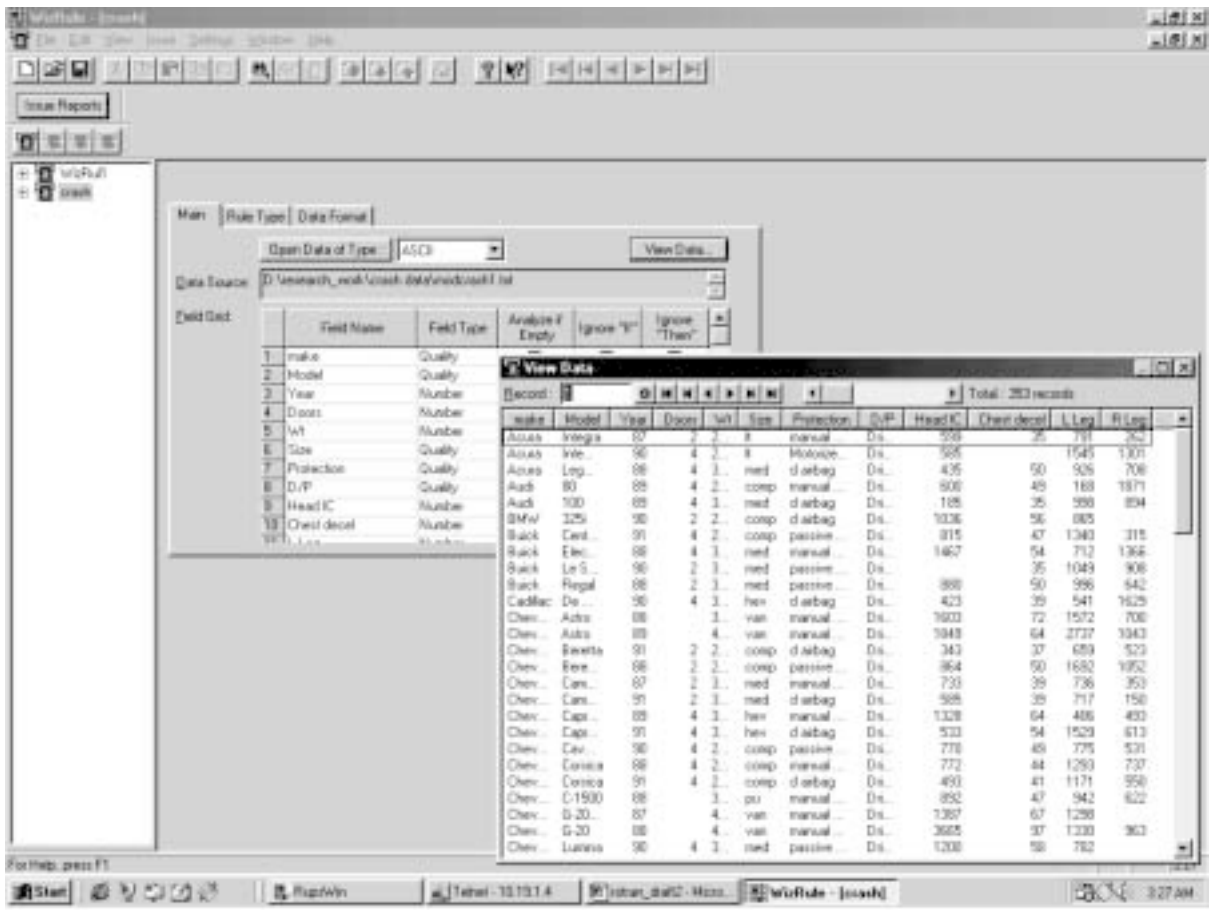

Fig 3. Typical WizRule user interface

Table 3. Typical rules of crash dataset

Rule for Head Injury Criterion	***Rule for Chest Deceleration***
If make is Chevrolet and Wt is 3,820.00 ... 5,619.00 (average = 4,514.40) and Protection is manual belts Then Head IC is more than 903.07 Rule's probability: 1.000 The rule exists in 10 records. *Significance Level: Error probability < 0.0001*	If make is Isuzu and Protection is manual belts Then Chest decel is more than 48.37 Rule's probability: 1.000 The rule exists in 14 records. *Significance Level: Error probability < 0.00001*
Rule for left femur load	***Rule for right femur load***
If Year *is* 90.00 and Protection *is* Motorized belts *Then* L Leg *is more than* 1,054.01 *Rule's probability:* 0.917 *The rule exists in* 11 *records*. *Significance Level: Error probability* < 0.001	*If* Doors is 4.00 and Protection is Motorized belts and D/P is Driver *Then* R Leg *is more than* 740.92 *Rule's probability:* 0.929 *The rule exists in* 13 records. *Significance Level*: *Error probability* < 0.0001

6. Comments and Future Research Directions

Data mining problem of vehicle crash data was demonstrated in a previous section. In the fields of transportation engineering problems discussed in the paper, there is further scope to study in depth on data mining applications. One needs to address problems in the context of data mining such as:

- Efficient approach for handling larger data bases with high dimensionality.
- Methods to overcome over-fitting of model
- Management of missing and noisy data
- Establishing complex relationship between fields
- Clear understandability of patterns
- Better user interaction and prior knowledge
- Possibility of Integration with other systems.

For detail discussion on above issues, readers may refer the reference [1].

7. Conclusions

Data mining is broadly defined as the search for interesting patterns from large amounts of data. Techniques for performing data mining come from a wide variety of disciplines including traditional statistics, machine learning, and information retrieval. While this means that for any given application there is probably some data mining techniques for finding interesting patterns, it also means there exists a confusing array of possible data mining tools and approaches for any given application. In this paper, the author has given brief background on Data mining and available commercially available tools. Various possible problems of transportation engineering have been discussed in the context of data mining. Finally, the applicability of commercially available data mining tool is used to solve the problem of 'crash data'. The results demonstrated that data mining could be an effective tool in the field of transportation engineering.

Acknowledgement

The author would like to thank **www.wizsoft.com** for allowing downloading their product to demonstrate the problem.

References

1. Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P. and Uthursamy, R. Advances in knowledge discovery and data mining, AAAI Press/The MIT Press, Cambridge, MA, 1996.
2. Kohavi, R. Data mining and visualization. In: Sixth Annual Symposium on Frontiers of Engineering, National Academy Press, D. C., 2001, p. 30–40.
3. Barai S. V. and Reich, Y. Data Mining of Experimental Data: Neural Networks Approach. In: Proceedings of 2nd International Conference on Theoretical, Applied Computational and Experimental Mechanics ICTACEM 2001, held during 27-30 December 2001, and organized by Department of Aerospace Engineering, Indian Institute of Technology, Kharagpur, 2001 (CD-ROM).
4. Amado, V. Expanding the Use of Pavement Management Data. In: Transportation Scholars Conference 2000, University of Missouri, 2000, www.ctre.iastate.edu/mtc/papers/amado.pdf
5. King, M. A.; Elder IV, J. F., and Abbott, D. W. Evaluation of Fourteen Desktop Data Mining Tools, In: IEEE International Conference on Systems, Man, and Cybernetics, San Diego, CA, October 12–14, 1998.
6. Elder IV, J. F., and Abbott, D. W. A Comparison of Leading Data Mining Tools. In: Fourth Annual Conference on Knowledge Discovery & Data Mining, New York, New York, August 28, 1998.
7. Abbott, D. W.; Matkovsky, I. P., and Elder IV, J. F. An Evaluation of High-end Data Mining Tools for Fraud Detection.

In: IEEE International Conference on Systems, Man, and Cybernetics, San Diego, CA, October 12–14, 1998.

8. Business. Data Mining Information at Business.com, 2002 http://www.business.com/directory/computers_and_software/software_applications/data_management/data_mining/index.asp
9. http://www.lascruces.com/~rfrye/complexica/dm_em.htm
10. http://www.mdx.ac.uk/www/roadtraffic/welcome.htm
11. http://soleunet.ijs.si/website/other/final_report/html/WP5-s5.html
12. Miller, H. J., and Shaw, S. GIS-T Data Models. In: Geographic Information Systems for Transportation: Principles and Applications, Oxford University Press, 2001 (ISBN 0195123948).
13. Seth, R.; Langley, P., and Wilson, C. Mining GPS Data to Augment Road Models. In: International Conference on Knowledge Discovery & Data Mining, San Diego, August 1999, p. 104–113.
14. http://web.mit.edu/cts/www/research/hancock.htm
15. http://www.romdas.com/surveys/sur-rgh.htm
16. http://lib.stat.cmu.edu/DASL/Datafiles/Crash.html
17. http://www.wizsoft.com