Data Sci. Eng. (2017) 2:224–231
https://doi.org/10.1007/s41019-017-0044-2

Using GUHA Data Mining Method in Analyzing Road Traffic Accidents Occurred in the Years 2004–2008 in Finland

Esko Turunen[1]

Received: 12 May 2017 / Revised: 4 August 2017 / Accepted: 6 August 2017 / Published online: 12 August 2017

Abstract The suitability of the GUHA data mining method in analyzing a big data matrix is studied in this report in general, and, in particular, a data matrix containing more than 80,000 road traffic accidents occurred in Finland in 2004–2008 is examined by LISp-Miner, a software implementation of GUHA. The general principles of GUHA are first outlined, and then, the road accident data is analyzed. As a result, more than 10,000 associations and dependencies, called hypothesis in the GUHA language, were found; some easily understandable of them are presented here. Our conclusion is that the GUHA method is useful, in particular when one wants to explore relatively small size, but still significant dependencies in a given large data matrix.

Keyword Data mining · Knowledge discovery in database · Traffic accident investigation

1 Introduction

The objective of this study is twofold: First we introduce the main features of the GUHA data mining method and its software implementation LISp-Miner, and then, we investigate by the GUHA method the 83,509 road traffic accidents that occurred in the years 2004–2008 in Finland and were registered by the police. These accidents resulted injuries in more than 20% of all the cases, and 1203 of them were fatal; in these, more than one human being was killed in 90 cases. By analyzing the road traffic accident data, it may be possible to find reasons and accident conditions that have not been known to traffic designers, legislators and other experts, and therefore have not been addressed. Moreover, choosing the road traffic accident data as the test material for the GUHA method we wanted to test the usability of the method for real and for everybody comparatively easy to understand data. The analyzed data has also been analyzed at Jyväskylä University by other data mining methods [1]; one objective of this study is to investigate whether the GUHA method also finds dependencies that are not reported in the study [1]. The structure of the paper is as follows. In Sect. 2, we give a brief description of data mining in general, and the GUHA method and related software LISp-Miner in particular. In Sects. 3 and 4, Finland's 2014–2018 road traffic accident data and related research questions are presented, and in Sect. 5, we introduce some interesting findings produced by the GUHA method. Section 6 contains summary and conclusion.

2 Data Mining, the GUHA Method and LISp-Miner Software

Data mining refers to computer-based methods to find from large data masses relevant dependencies and features that are useful for the owner of the data. Typically, the data analyzed by data mining methods is, for example, measurements of the industrial process, extracts from the customer database or as in this study, systematically reported data on the road traffic accidents. In most cases, the computation algorithms used in data mining include various kinds of clustering methods, correlations, neural networks, self-organizing maps. In the successful exploitation of data mining, the most essential is the comprehensive

✉ Esko Turunen
esko.turunen@tut.fi

[1] Tampere University of Technology, Tampere, Finland

understanding of data and its various quantities. Even a mere innovative approach to data visualization, for example, can help to see the benefits of the given data from a completely new perspective.

The GUHA method, introduced in 1966 (see [2]), is one of the oldest data mining methods. GUHA is an abbreviation for General Unary Hypoteses Automation. According to its name, the aim in the GUHA method is not to test hypotheses on a given data, but rather to produce them under certain logical principles and user-defined, guiding guidelines for possible further research. GUHA can be used to study large data matrices. In practice, by presently available software, the examined data matrix may have dozens or a few hundreds of columns and hundreds of thousands of rows, and any single cell in the data matrix may contain any symbols. There is no need to assume any possible statistical distribution on the data. The strength of the GUHA method is its logical basis, the formalism to express statements in first-order monadic logic with generalized quantifiers and truth and falsehood of the sentences being defined in finite models.

An essential part of the GUHA logic language is *generalized quantifiers*; they can be used to find answers to questions about data containing expressions like 'almost all,' 'most,' 'significantly different subset,' 'quasi-equivalent subsets,' etc. So, say, 'Does the road accident data contain information on some special circumstances in which there has been exceptionally many fatal accidents?' or 'Is heavy vehicle involvement in the accident significantly dependent on some specific factors?'. With GUHA method, it is possible to find all the dependencies that the data supports; thus, it allows the user to get answers to the question 'Does my data contain some information not even occurred to me to look for?'.

The GUHA method was initially developed as a purely theoretical concept, but later it has been implemented in different kinds of computer software, from which the most extensive and significant is the LISp-Miner software developed at the University of Economics, Prague since 1996, and is freely down loadable [3]. LISp-Miner software is based on number crunching; in theory, the software tests all the possible alternatives that can be billions depending on the size of the data and the complexity of the task, but in practice—thanks to the logical criteria—only a few proms of all options have to be really computed. To our knowledge, general computational complexity issues of the GUHA method are, however, unsolved research problems.

The results generated by the LISp-Miner software are mostly *local* in nature, with the example of 'Collision with a deer occurs to middle-aged men significantly more frequent in the early afternoon than before noon' rather than *global* like 'The proportion of road entrenchments decreases linearly as the driver's age increases.'

Unlike Bayesian reasoning approach [4] or a neural network-based data mining [5], the GUHA method is not a black box method; the user must have some kind of preconception of the data being extracted. The use of the LISp-Miner software requires pre-processing of the data to be extracted prior to mining; for example, the data must be victimized, the user decides how small subdivisions of data are still worth exploring, how variables are divided into categories, and what kind of generalized quantifiers should be used; the LISp-Miner software has dozens of opportunities, for example, to model expressions like '*most*,' '*often follows*,' '*almost equivalent sets*,' '*much larger than the average*.'

The key part of the GUHA method is *analytical questions*. They are natural language expressions related to the data being studied. Examples of analytical questions related to the road traffic accident data are 'How do accidents involving a young driver differ from other accidents?' or 'What is typical for accidents involving a heavy vehicle?'. Analytical questions can be formalized in the GUHA language and then analyzed with LISp-Miner software. Formally, GUHA logic formulas are of the form $\phi \approx \psi$, where ϕ and ψ are logic combinations of variables of the data (corresponding to the columns of the data matrix) and $\approx$ is a generalized quantifier. In practice, the LISp-Miner software generates fourfold contingency tables that are then examined in the light of the user's criteria. A fourfold contingency table has the form:

	ψ	$\neg\psi$	
ϕ	a	b	$a+b=r$
$\neg\phi$	c	d	$c+d=s$
	$a+c=k$	$b+d=l$	m

where $a+b+c+d=m$, the number of rows in the data matrix, and

- a is the number of objects satisfying both ϕ and ψ,
- b is the number of objects satisfying ϕ but not ψ,
- c is the number of objects not satisfying ϕ but satisfying ψ,
- d is the number of objects not satisfying ϕ nor ψ.

At present there are 21 quantifiers implemented to LISp-Miner [6]. In general, there are various types of generalized quantifiers in GUHA theory formalizing various kinds of associations:

- *Implicational Quantifiers* formalize the association *Many* ϕ *are* ψ, they do not depend on the values c, d.
- *Comparative Quantifiers* formalize the association ϕ *makes* ψ *more likely than* $\neg\psi$.

- Some quantifiers just express observations on the data, and some others serve as tests of statistical hypotheses on unknown probabilities.
- Some quantifiers are *symmetric*: $\phi \approx \psi$ implies $\psi \approx \phi$, and some *admit negation*: $\phi \approx \psi$ implies $\neg\phi \approx \neg\psi$.

An advantage of GUHA is that new quantifiers can be defined and their properties can be analyzed in a well-established logic framework. LISp-Miner software has currently six different computing procedures. We used three of them in this study: 4ft-Miner, SD4ft-Miner and Ac4ft-Miner. 4ft-Miner procedure analyzes two formulas, the Antecedent (denoted by ϕ) and the Succedent (denoted by ψ); as said they are combinations of columns of the data matrix. If the given truth criteria are met, by logic terminology $v(\phi(x) \approx \psi(x)) = \texttt{TRUE}$, the software prints the result as a found *hypothesis*. If the criteria are not met, the phrase is false and will not be printed. Thus, the truth TRUE and falsehood FALSE of a logic formula $\phi \approx \psi$ is based on the values in the fourfold contingency table. For example, if $\approx$ is a *founded p-implication quantifier*, where $p \in (0, 1]$ and denoted by $\Rightarrow_{p,\text{Base}}$, we define $v(\phi(x) \Rightarrow_{p,\text{Base}} \psi(x)) = \texttt{TRUE}$ if $\frac{a}{a+b} \geq p$ and $a \geq \text{Base}$ and $v(\phi(x) \Rightarrow_{p,\text{Base}\psi(x)}) = \texttt{FALSE}$ elsewhere. If

$$v(\phi(x) \Rightarrow_{p,\text{Base}} \psi(x)) = \texttt{TRUE},$$

then at least $p\%$ of objects (rows in the data matrix) satisfying ϕ satisfy also ψ, in other words, founded p-implication quantifiers are related to the truth of the statement

$\phi(x)$ implies $\psi(x)$ with confidence p and support Base.

Another example is *Fisher quantifiers* corresponding to the test of hypothesis

$$\text{Probability}(\phi(x)|\psi(x)) > \text{Probability}(\phi(x)|\neg\psi(x))$$
$$\text{with significance } \alpha$$

is defined to be TRUE in the data (or supported in the data) if $ad > bc$ and

$$\sum_{i=0}^{\min\{b,c\}} \frac{r!s!k!l!}{m!(a+1)!(b-1)!(c-1)!(d+1)!} \leq \alpha.$$

and FALSE elsewhere. Useful quantifiers are also the *Above average quantifiers* $\sim_q^+$, where $q \geq 0$; if $v(\phi \sim_q^+ \psi) = \texttt{TRUE}$ then among objects (rows in the data matrix) satisfying ϕ there are at least $q\%$ more objects satisfying ψ than in the whole data matrix. In other words, the presence of ϕ makes the presence of ψ at least $1 + q$ times more frequent. Obviously,

$$v(\phi \sim_q^+ \psi) = \texttt{TRUE} \text{ iff } \frac{a}{a+b} \geq (1+q)\frac{a+c}{a+b+c+d}.$$

These quantifiers and many others are implemented in 4ft-Miner, a procedure of LISp-Miner. More diverse and complex data mining tasks can be carried out by SD4ft-Miner and Ac4ft-Miner, the other procedures of LISp-Miner. SD4ft-Miner and Ac4ft-Miner procedures analyze four formulas. The truth and falsehood of associations corresponding to the related quantifiers are based on two contingency tables

	ψ_1	$\neg\psi_1$	
ϕ_1	a_1	b_1	$a_1 + b_1 = r_1$
$\neg\phi_1$	c_1	d_1	$c_1 + d_1 = s_1$
	$a_1 + c_1 = k_1$	$b_1 + d_1 = l_1$	m

	ψ_2	$\neg\psi_2$	
ϕ_2	a_2	b_2	$a_2 + b_2 = r_2$
$\neg\phi_2$	c_2	d_2	$c_2 + d_2 = s_2$
	$a_2 + c_2 = k_2$	$b_2 + d_2 = l_2$	m

As an example, by

$$\left|\frac{a_1}{a_1+b_1} - \frac{a_2}{a_2+b_2}\right| \geq p \in (0,1], a_1 \geq \text{Base}_1 \text{ and } a_2 \geq \text{Base}_2$$

we define the truth that two subsets in the data matrix differ considerably from each other with respect to some property.

By the following two examples, we briefly introduce the use of the 4ft-Miner and Ac4ft-Miners software, more detailed information can be found, e.g., in [7]. An analytical question 'Are some types of accidents particularly typical for female drivers?' can be analyze with the 4ft-Miner procedure through the following steps

1. Set on the Antecedent part ϕ all the variables (corresponding data matrix's columns) whose impact on accidents is to be investigated. In addition, attach certain parameters, e.g., how many variables are to be included, whether they associated with an 'and' or 'or' -connective, etc.
2. Set as Succedent ψ the variable 'Female driver,' which corresponds to one of the data columns, and retain the desired parameters.
3. Decide how small subsets are still considered relevant; the *Base* is in this case it is fixed to 100.
4. Choose the generalized quantifier; here it is *founded p-implication* with $p = 0.6$. This means that if there is a sufficiently large accumulation of accidents (i.e., $a \geq 100$) involving at least 60% of a female drivers;

$\frac{a}{a+b} \geq 0.6$, then LISp-Miner prints this result as a hypothesis.

Then computing can be started. In this example, 4ft-Miner goes through more than 45 million fourfold contingency tables and produces four cases where the criteria are met. On of them is the following hypothesis 'The data includes 166 accidents on icy road that occurred in the morning or day time and involved a 36- to 55-year-old driver. 102 drivers out of 166 were women.' Given that women were involved in about 25.5% of all over 80,000 road accidents, but in this case 61.5%, it should be assumed that even more experienced female drivers are more vulnerable to serious accidents in daylight on frozen roads. In general, it is up to the user to evaluate whether the results are significant or not. Moreover, there is also a module in 4ft-Miner that tests afterward in Bayesian sense the probability of the truth of hypothesis produced by 4ft-Miner software (see [8]).

More complex than the 4ft-Miner procedure is the SD4ft-Miner procedure (SD is an abbreviation of *set differs from set*), which searches for distinct subdivisions of the data, and the Ac4ft-Miner procedure, which searches for subdivisions that are similar to some variables, but are clearly different for some other variables (Ac in an abbreviation of *action*). These data mining tools are based on a comparison of two fourfold contingency tables with respect to user-defined parameters. As an example, assume we want to investigate whether there are some types of accidents involving drivers of different ages that different in frequency at least 25%. The analytical question would then be 'Are there some types of accidents that are much more frequent is some age group than in some other age group?'. After setting the parameters and running Ac4ft-Miner, we obtain several hypothesis; one of the results is 'For drivers under 25 years of age, there were far more derailing from the road accidents in winter on certain types of roads than in comparable conditions for 45–64-year-old drivers.' In these circumstances, a total of 3988 accidents occurred for young drivers, of which 2298 (58%) were derailing from the road accidents, while the corresponding figures for older drivers were 4071 and 1226 (30%). Therefore, it can be said that in these circumstances, derailing from the road accidents was more typical for younger than old drivers.

3 Research Data and its Variables

The survey data was originally collected from the Registry of the Finnish Police and the Finnish Transport Agency. The original data has been modified at Tampere University of Technology and University of Jyväskylä to fit for data mining tasks; for example, age has originally 94 categories (age in years 5,...,98), and here they are connected to only 7 categories. The list in Table 1 is only indicative. The meaning of some variables might remain obscured; however, this is irrelevant because the main aim is to introduce the GUHA method and not the traffic accident study in detail. It is enough to know that there are 83,509 rows in our data matrix; each of them corresponds to one reported road traffic accident that occurred in the years 2004–2008 in Finland, and more than 100 columns corresponding to information collected by the road authorities.

All computer runs made in this study were performed using a computer system called Techila PC-Grid [9]; at TUT there are about 400 desktops computers that have been interconnected in a way that allows computing time demanding tasks to be computed 10–100 times faster than comparing with one only table computer. Typical road traffic accident data mining LISp-Miner running time was from a few minutes to several hours. With one single computer, the longest runs would have lasted for days, and all in all, computer runs would have lasted for 300 days; by Techila PC-Grid they were done in a few weeks. During that time, Techila PC-Grid checked up billions of fourfold contingency tables. Primary, to keep thinks simple, we wanted to find as simple associations as possible.

4 Analytical Questions Asked in the Study

In this study, our aim was to respond to the analytical questions listed below by using the GUHA data mining method. The answers to some of them are commonly known; for example, a large part of accidents caused by drunken drivers are known to be driving out the road and other one vehicle accidents. It is also known that head-on collisions at high speed with a heavy vehicle often lead to serious consequences. Of course, our goal was to find new, possibly surprising results. We presented and implemented the following analytical questions for the LISp-Miner software.

1. Do some specific factors anticipate certain types of accidents, that is, do certain types of conditions imply particularly one vehicle accidents, pedestrian accidents, collision with an animal, etc?
2. Which factors predict accidents leading to personal injury?
3. Property damage in accidents. What kind of accidents resulted other property damage than damage to vehicles (such as a broken traffic sign, light bulb, a railing, as a result of an accident)?
4. Accidents caused by drivers of different age, differences between different age groups. (a) Were there accidents that happened particularly often in a certain

Table 1 Data variables and the number of their categories. The above mentioned variables were mainly used for data mining tasks

VARIABLE	ABBREVIAT	CATEGORIES, pcs, names
Drivers		
Drug / medicine	drug-pill	5 pcs
Age	Age	7 pcs
Driving license	Licence	4 (yes, no, new driver, foreigner)
Alcohol	Alco	2 (yes, no)
Alcohol per mil	Premil	8 pcs
Gender	Gender	2 (female, male)
Conditions		
Temperature	Temp	3 (frost, close to zero, above zero)
Road surface	Surface	7 pcs
Weather	Weather	6 pcs
Road work	Road Work	2 (yes, no)
Lightness conditions	Light	4 pcs
Accident and consequences		
Dead	Dead	2 (yes, no)
Injured	Injured	2 (yes, no)
Deads	Deads	4 (0, 1, 2 or more)
Many injured	ManyInj	4 (0, 1, 2, 3 or more)
Other property damage	OtherDama	5 pcs
Accident category	AcCateg	13 pcs
Accident type	TypeAccid	55 pcs
Number of casualty	CasualNum	4 (1, 2, 3, 4 or more)
Heavy vehicle	Heavy	2 (yes, no)
Place and time		
Month	Month	4 (spring, summer, fall, winter)
County	County	5 pcs
Accident location	Location	8 pcs
Road maintenance area	RoadMain	9 pcs
Hour	Hour	8 pcs
Weekday	Day	7 pcs
Intersection		
Traffic lights	Traflights	4 (in operation, flashing, inactive, inoperative)
Crossroads	CrosswayVar	8 pcs
Road entry class	RoadEnt	4 pcs
Junction type	JunctionType	6 pcs
Other interface	OtherInt	7 pcs
Road and its condition		
Number of runways	RunNumber	4 pcs
Housing	Housing	6 pcs
Winter maintenance class	WinClass	6 pcs
Average daily traffic	ADT	5 pcs
Highway, motor (traffic) road	MoTR	3 pcs
View percent 300 m	View300	5 pcs
Speed limit type	Speed LT	6 pcs
Speed limit	SpeedLimit	7 pcs
Road coating	Coating	4 pcs
Road width	RoWidth	6 pcs
Road category	RoadCat	4 pcs
Walkway	Walkway	2 (yes, no)
Heavy truck average daily traffic	HeavyTruck	5 pcs
Village	Village	2 (yes, no)

age group? (b) What were the differences in accident types between drivers of different ages? Were some types of accidents much more typical for young drivers than for elder drivers, or the opposite?

5. Typical accidents of new drivers, drivers without driving license and foreigners. Which types of accident are typical accidents for new drivers or drivers without a valid driving license? Do foreign heavy vehicle drivers cause particular road accidents?
6. Accidents in the areas of work on road. What kind of accident was a major part of certain kind of accidents in the road work area? Did a large part of certain accidents happen specifically in the area of work on road?
7. Typical accidents of men and women: the differences between the sexes. (a) Was there a major part of certain types of accidents for men or women only? What were the typical road accidents of men or women? (b) Are certain road accidents typical for women than men, or the opposite?
8. Accidents at roundabouts. (a) Was there a large part of a certain type of accident that occurred in roundabouts? Was there a large part of particular crossroad accidents that happened in particularly at roundabouts? (b) Did traffic accidents occur more frequently or less frequently at roundabouts than in other crossroads?
9. Regional differences. Was there any difference with respect to accidents leading to injuries between different parts of Finland? Were certain types of accidents typical to particular parts of Finland?
10. Impact of alcohol, drugs and medicines in road accidents, and differences between them. (a) What kind of accidents were typical of drivers under the influence of alcohol or drugs? (b) How did typical accidents that occurred to drivers under the influence of alcohol differed from accidents that occurred to sober drivers? (c) Did drivers who had used drugs have similar accidents than divers who had been drunk?
11. The impact of speed limitations on road traffic accidents. Did speed limitations have a correlation to the number of injured or deceased in accidents? Did speed restrictions effect on types of accidents?
12. Accidents at different times of year and day. Was any accident type dependent on the time of the day or the season of the year? Was there a difference between other vice similar accidents that caused injured or deceased but differed only in terms of time of the day or season of the year?
13. The impact of circumstances (such as weather, road pavement and surface, brightness, road maintenance and congestion) on road accidents? What kind of accidents were typical in rainy weather or winter conditions, or on a dry road?
14. The impact of heavy vehicles. Did the involvement of a heavy vehicle effect on the accident type? Was there such types of accidents that are typical for a heavy vehicle?
15. Collision accidents. It is commonly known that collision accidents often result personal injuries, but what other factors are involved in collision accidents? Were the collision accidents in some circumstances particularly dangerous? Where and when did collision accidents happened a lot?

5 Some Interesting Results Found Out by the GUHA Method

The GUHA analysis was carried out in 2012, and the results were published first in the [10]; a total of more than 10,000 hypotheses were found and 100 most interesting ones were reported. Here we present some of them. The purpose is not to explain in detail how the results were found, but to give an idea of what kind of results were found. Many of the discovered dependencies are certainly known facts for traffic investigators. This confirms that the GUHA method can be used to find meaningful dependencies on the given data. We mention here some of them. However, the most meaningful are the findings that are unknown even to the experts in the field; we give some examples of these, too.

1. *One Vehicle Accidents* (that is, there were no other involved parties) was the largest accident category, 24,803 cases (29.7%) of all the accidents. In this category, driving under the influence of alcohol was a major security risk. This well-known fact was reflected also in this study in many ways. For example, the data contains 1369 one vehicle accidents under the influence of alcohol, of which 1005 (74.3%) led to injuries. Also drivers below the age of 36 were at night and early in the morning a high safety risk for themselves and their passengers; accidents leading to serious injuries that occurred outside a junction during the night time totaled 1334, of which 1006 (75.4%) were one vehicle accidents.
2. There where 4089 *collision accidents* of two vehicles, or 4.97% of all the accidents. They related to the following dependencies. Colliding with a heavy vehicle was very often fatal. This expected result is confirmed by this data. Of all the fatal 284 accidents occurred outside an intersection, there where 234 cases (82.4%) in which a heavy vehicle was involved. Moreover, particularly on a straight dry road with road speed limitation, 110 head-on collision with heavy vehicles occurred: 52 of these (47.3%) resulted in

death. This raises suspicion of driver's suicide, falling asleep, etc. The truth of the matter cannot be deduced from the data.

3. *Accidents that Involved Pedestrians* were proportionally the most fatal of all the accidents that caused injuries: 625 cases (80%) of all 779 pedestrian accidents caused injuries or deaths. Typically, the driver of a vehicle that ran on a pedestrian had a valid driver's license. 560 (72%) of such an accident took place outside pedestrian crossing.
4. 1389 of all the 1549 *bicycle accidents* were collisions with a vehicle, of which 985 (66%) caused one person to be injured.
5. *Moped Accidents* were often fatal to young drivers: under the age of 25 were injured 507 (56.4%) of all the 899 persons involved in such accidents, also 88 (48%) out of the 181 accidents, 25–34-year-old moped drivers were injured. The result is partly explained by the fact that a typical motorized vehicle of young person in Finland is moped.
6. *Animal Accidents*, i.e., accidents in which a vehicle crashes into a wild animal usually took place at night. Out of 13952 registered animal accidents 10,950 (78%) occurred in dark time, mostly in fall. Other typical features of animal accidents were sparsely populated area; however, the driver was mostly not under the influence of alcohol.
7. In particular, *a collision of a vehicle with a deer (not moose)* almost always avoided personal loss. These accidents were recorded 14,435, and only two of them resulted in human's death.
8. The number of injure resulting one vehicle accidents caused by *young drivers* (in this case under 35 years of age) occurred relatively much more in the evening than in the morning or at noon. Between 10 and 15 h, these accidents accounted for 33% (155 cases out of a total of 472 cases), while between 18 and 21 h 65% (108 cases out of a total of 165 cases). Most of this type of accidents occurred in the summer time on gravel roads. For drivers under the age of 25 who had a driving license, there were more injuries in crossings in the summer time than in the corresponding circumstances in winter. These accidents occurred in bright weather. In summer, accidents without injuries or deaths accounted for 52% (280 out of 538 accidents), while the percentage in winter was higher, 78% (162 out of 207 accidents). In summer, the absolute number of occurred injuries was much bigger than in winter (258 injuries in summer, but in winter only 45 accidents resulted personal injury).
9. *Gender Impact on Traffic Accidents* Accidents involving men have been reported in 59,429 (71.2%) cases and women 20,940 (25.1%) cases, data was missing in 3140 (3.8%) cases. Given the lack of data on what is the proportion of male and female drivers in all road traffic in Finland, it cannot be concluded whether the one sex has a higher probability for accidents than the other. However, about men and women who have suffered accidents can be said something; for example, the following associations were found. Accidents that occurred in bypassing situation and involving a heavy vehicle happened to women in relative terms much more often than for men. For example, in good-condition roads outside the intersection area, a total of 2999 accidents were reported to men and 696 (23%) of these accidents were bypass accidents. As many as 202 (52%) of the 418 accidents pertaining to women in similar conditions were bypass accidents. Another typical accident for women at 90–100 km/h speed limit area in the late evening was an animal accident. Of all the 437 reported accidents involved a female driver 331 (76%) were a collision with an animal. In most cases, the road surface was dry, the accident occurred in a sparsely populated area and the driver was 45–64 years old. On the other hand, almost always of the accidents that occurred at night or early in the morning, involved a male driver. Between midnight and at 6 o'clock, 12,649 accidents occurred, with a male driver in 10,409 (82%) cases. However, very far-reaching conclusions can not be made; we do not know what were the proportions of male and female drivers in road traffic at night.
10. *Regional Differences in Traffic Accidents* Moose collisions occurred in sparsely populated areas were, relatively speaking, much more frequent in the Province of Oulu than in the Province of Southern Finland and the province of Western Finland. In the province of Oulu, 1546 out of the 2419 in all accidents (45%) in sparsely populated areas were moose collisions, whereas in the Province of Southern Finland the corresponding figures were 1192 and 6349 (19%) and in Western Finland 2345 and 11,966 (20%), respectively. Typically, the driver was male.

6 Summary and Conclusion

In this study, we investigated the suitability of the GUHA data mining method and the LISp-Miner software to extract relevant information from a big data matrix in general, and in particular a $100 \times 83{,}500$ data matrix related to road traffic accidents that occurred in Finland in 2004–2008. GUHA is a first-order monadic logic-based data mining method; the main concepts are generalized quantifiers and analytical questions set by the user; the analytical questions can be expressed by GUHA language. The truth or

falsehood of a single association or relation between two logic expressions in a data is verified by one or two contingency tables. The strength of the GUHA method is a robust logical and mathematical basis. New ideas and ways of extracting information from large data matrices can be defined in the GUHA language, explore their features, and ultimately implement them in software. GUHA has evolved over 50 years and is constantly evolving. However, a practical data miner does not need to master all the theoretical basis of GUHA in-depth. LISp-Miner is a software [3] implementing GUHA; according to the GUHA principles, LISp-Miner goes through contingency tables and verifies or falsifies relation between logic expressions, Boolean in nature. More precisely, the patterns the LISp-Miner deals with contain several derived Boolean attributes (from two in the 4ft-Miner to five in the SD4ft-Miner). Moreover, the procedures described in Section 14.9 in [6] deal also with categorical attributes not transformed into Boolean ones. However, we have not used these latter procedures in this study.

The road traffic accident data matrix covered more than 83,500 rows (each corresponding to a road traffic accident) and more than 100 columns (factors related to the accidents). A single PC would have needed over 300 days to compute all the required billions of contingency tables, but utilizing TUT's grid system, which networks up to about 400 desktop computers at TUT campus [9], computation took only a fraction of that time. In total, the computation yielded more than 10,000 associations called hypotheses in GUHA language, of which 10 easily comprehensible are reported in this report.

The GUHA method produces hypotheses such as 'The risk of colliding with a moose by car (a very typical road traffic accident type in Finland)in the Oulu region is triple compared to the risk in the Häme region.' However, the GUHA method does not explain the deeper connections behind the hypotheses; there might be relatively more moose in the Oulu region than in the Häme region. All this depends on the investigator's interpretation and may require further studies. Naturally, GUHA also produces trivial, well-known results. Some of these are included in this report as it is intended to demonstrate that if there is any significant dependence on the data, LISp-Miner software finds it.

As a result of the study, we conclude that the GUHA method and the LISp-Miner software can be used to extract from a road traffic accident type data dependencies that cannot be found by other data mining methods. For example, in a study conducted at the University of Jyväskylä, no such detailed dependencies were found on the same data as this report presents [1]. On the other hand, this is understandable since GUHA is a unique method and its findings cannot be directly compared to those of other data mining methods. Although the current interface of the LISp-Miner software is a bit embossed, it can be recommended in analyzing large data alongside other data mining algorithms, especially when it comes to studying dependencies between relatively small, but still significant subsets of a big data repository.

References

1. Äyrämö S, Pirtala P, Kauttonen J, Naveed K, Kärkkäinen T (2009) Mining road traffic accidents. University of Jyväskylä, Jyväskylä
2. Hájek P, Havránek T Mechanizing Hypothesis Formation. http://www.cs.cas.cz/hajek/guhabook/
3. Rauch J, Simunek M LISp–Miner home page. http://lispminer.vse.cz/
4. B–course home page. http://b-course.cs.helsinki.fi/obc/
5. Hand DJ, Mannila H, Smyth P (2001) Principles of data mining. The MIT Press, Cambridge
6. Rauch J (2013) Observational Calculi and Association Rules. Studies in Computational Intelligence. Vol. 469. Springer, Heidelberg
7. Turunen E Lecture notes. http://math.tut.fi/~eturunen/AppliedLogics2008.html
8. Piché R, Järvenpää M, Turunen E et al (2014) Bayesian analysis of GUHA hypotheses. J Intell Inf Syst. doi:10.1007/s10844-013-0255-6
9. Techila home page (in Finnish). http://www.techila.fi/
10. Järvenpää M, Turunen E (2012). Suomessa 2004–2008 sattuneiden tieliikenneonnettomuuksien analysointia GUHA–tiedonlouhintamenetelmällä (in Finnish), Tampere University of Technology. Department of Mathematics. Research Report, Vol 99