A DATA QUALITY CASE STUDY FOR TURKISH HIGHWAY ACCIDENT DATA SETS

Rahime Belen[1], Tuğba Taşkaya Temizel[1], Ömür Kaygısız[2]
[1]METU Informatics Institute, Ankara, Turkey
[2]The General Directorate of Security, Traffic Research Center
E-mail: rbelen@ii.metu.edu.tr, tugbatt@ii.metu.edu.tr, okaygisiz@egm.gov.tr

ABSTRACT

The highway accident data provides valuable information such as accident black spot locations and their spatio-temporal change to the experts. With the help of this information, experts may take timely precautions in order to prevent the incidents from happening in the future. There is a challenging detail here. Highway accident data is often manually collected. For example, an officer submits the whereabouts of accidents in terms of latitude and longitude on a form or on a map to the system. Although many systems are designed to verify the legitimacy of the data to be entered, systems can still be prone to user errors. Users may enter illegitimate values which appear to be valid in the system, i.e. a point that is not on a highway (an outlier) or a point on the highway but not correctly entered (an inlier). These erroneous values are called *disguised missing data* and yet can arise in many different data sets such as health or survey apart from spatial data sets. However, the way they emerge and the way they appear in spatial data sets such as in the highway accident data sets is different and detection is difficult compared to that of survey and health-related data sets. Consequently, their presence can affect the outcome of data analysis tasks severely which may cause decision makers to make inaccurate decisions. Therefore elimination of these values becomes a necessity prior to the data analysis.

In this paper, we will explain the common disguised missing value problems in Turkish highway accident data sets. Since the hotspot analysis will be run after data quality is guaranteed, we have focused on disguised missing values on coordinate information. We will describe a framework about how to detect such values automatically. We believe that the results of this study will be of benefit to the data analysts who are working with similar data sets.

1 INTRODUCTION

Every day important decisions are made based on the data stored within databases. There is a challenging detail; the decisions that are made are only as good as the data upon which they are based on (Hulse, Khoshgoftaar, & Huang, 2007). Consequently data quality has a severe impact on analysis.

There are many definitions of data quality but the most frequently used one is: A collection of data X is of higher quality than a collection of data Y if X meets customer needs better than Y (Redman, 1997). Most basically it is the fitness of use which implies that data quality is inherently subjective.

To Cite This Article: *Belen, R., Tuğba, T., ve Kaygısız, Ö., "A Data Quality Case Study for Turkish Highway Accident Data Sets", 15. International Conference Road Safety on Four Continents, UAE, Abu Dabi, s.344-456, 2010*

Data quality is a multidimensional concept. In data quality assessments each organization must determine which dimensions are important to its operations and precisely define the variables that constitute the dimensions.

The accuracy of data can be affected negatively by disguised missing data which is the main topic of this paper. Some user interface or database designs may lead user to enter spurious values. Databases that are designed so as not to allow any "null" or "unknown" values can be given as an example. In such cases, users may act in two ways: The first one is to choose the first available entry in a select box (if provided). For example attribute birth date is generally wanted to be disclosed and January 1 (the first value in the pop-up lists of month and day, respectively) may be chosen in order to skip the question.

The second one is to enter any legitimate value to indicate "unknown". For example, consider a system where the user is required to select a certain city as place of birth but only the cities in Turkey can be selected from the list. If the user was born outside Turkey, the user will tend to select the first available entry in the list such as city of "Adana". On the other hand, there will be people who were actually born in Adana. As a result, the system will comprise both accurate and inaccurate data entries. If a legitimate value is used frequently in lieu of missing values, it is called a disguised missing value.

2 BACKGROUND

While formulating the strategies and safety plans or assessing the effectiveness of road safety programs, the availability of accurate and reliable data is of paramount importance. However, information systems that are designed to collect road safety data may be implemented without taking into account data collection and quality factors. As a consequence, to derive statistics from these databases individually as well as collectively often constitutes a handicap. The reason of poor quality data is that data normally does not originate from systems that were set up with the primary goal of mining this data.

For example, it is difficult to compare the statistics derived from such road accident databases in different countries due to the discrepancies in the data collection policies. The varying definitional issues such as "road safety fatality" in different countries, underreporting problems and inaccurately collected data are the most common data quality problems encountered in road safety databases (Peden & Toroyan, 2005). Although there is an international agreement on the definition of "road safety fatality" as "any person killed immediately or dying within 30 days as a result of an injury accident", many countries may apply different definitions. Underreporting problems may occur but can be tackled to some degree by utilizing the output of hospital information systems together with the police databases. A lack of professionals trained in road safety means that data such as the number of people seriously injured, crash location, type of road user (pedestrian or not), and fatalities may not be collected correctly thus impairing the quality of data required to evaluate road safety interventions.

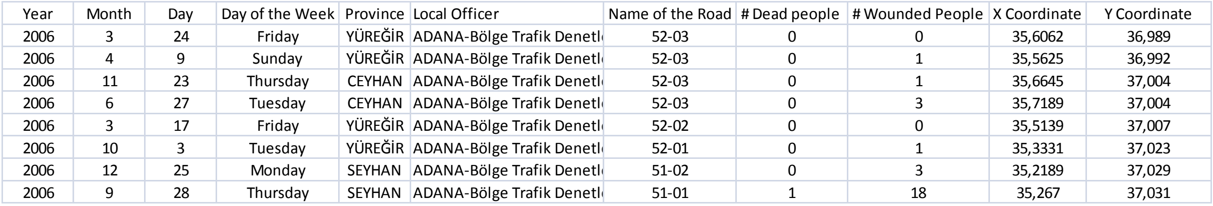
Many studies have been conducted to assess the data quality of road safety databases. The three aspects of data quality that are completeness, timeliness and accuracy have been measured in Canada's National Collision Database (Chouinard & Lecuyer, 2009). They have found that while database completeness differs significantly among the province/territories for fatal and serious injury collisions, there appears to be a tradeoff between the timeliness and completeness of data in many jurisdictions.

The completeness factor has also been measured for road traffic crash data in Ghana (Salifu & Ackaah, 2009). The level of under-reporting has been analyzed by generating relevant alternative data based on surveys conducted at hospitals and among drivers and then matching this data with that of records in police crash data files and official databases. Such a data linkage methodology is one of the most common approach in which routinely collected police reports of traffic accidents and hospital discharge files are individually matched or "linked" using a computerized procedure. Using additional data resources in order to improve data quality and to obtain more accurate statistics is a widely used technique. For example, a risk index has been developed which is not solely based on collision statistics but also subjective evaluation techniques (Leur & Sayed, 2002).

Advancements in technology such as the use of GPS devices may help to improve the data quality but are not yet alone sufficient. A good data quality policy should also take into consideration the human factors, i.e. competency in accident investigation, attention in reporting, constant user feedback to induce system improvement are just few to name (Transportation Research Board National Research Council, 1991). Standard definitions, implementing quality control systems, simplifying data requirements, and mandatory reporting are also accounted for part of a successful data quality plan (Elsig, 2009).

3 DATABASE CHARACTERISTICS

Road safety accidents are stored in a relational database which is used by all traffic inspectorates in Turkey. When an accident takes place, the local officer should enter the details of the accident, i.e. time and date of the accident, town, city, the road name, road type, accident type, number of dead and wounded people, latitude and longitude of the accident. All of which are mandatory. You can see a subset of accidents recorded in Adana in Table 1.



Year	Month	Day	Day of the Week	Province	Local Officer	Name of the Road	# Dead people	# Wounded People	X Coordinate	Y Coordinate
2006	3	24	Friday	YÜREĞİR	ADANA-Bölge Trafik Denetl	52-03	0	0	35,6062	36,989
2006	4	9	Sunday	YÜREĞİR	ADANA-Bölge Trafik Denetl	52-03	0	1	35,5625	36,992
2006	11	23	Thursday	CEYHAN	ADANA-Bölge Trafik Denetl	52-03	0	1	35,6645	37,004
2006	6	27	Tuesday	CEYHAN	ADANA-Bölge Trafik Denetl	52-03	0	3	35,7189	37,004
2006	3	17	Friday	YÜREĞİR	ADANA-Bölge Trafik Denetl	52-02	0	0	35,5139	37,007
2006	10	3	Tuesday	YÜREĞİR	ADANA-Bölge Trafik Denetl	52-01	0	1	35,3331	37,023
2006	12	25	Monday	SEYHAN	ADANA-Bölge Trafik Denetl	51-02	0	3	35,2189	37,029
2006	9	28	Thursday	SEYHAN	ADANA-Bölge Trafik Denetl	51-01	1	18	35,267	37,031

Table 1: Highway Accidents Database show the accidents information for provinces in Adana city

The system implements basic data quality checks. For example, an officer is not allowed to enter a coordinate outside the region of the traffic inspectorate. The system lets officers to put the requested information only in specific formats and range such that users cannot select an invalid date or road name.

4 A DATA QUALITY PROBLEM: DISGUISED MISSING DATA

Although the system does some checks with respect to ensuring data integrity and quality, it still suffers from some problems. The typical problems we have encountered are as follows:

a. GPS devices may malfunction due to a problem in the projection system and consequently inaccurate coordinate information may be obtained.
b. Longitude information may be entered in lieu of latitudes and vice versa. This situation arises when these coordinates are very close in a particular town.
c. Users may enter the coordinate information inaccurately in the database.

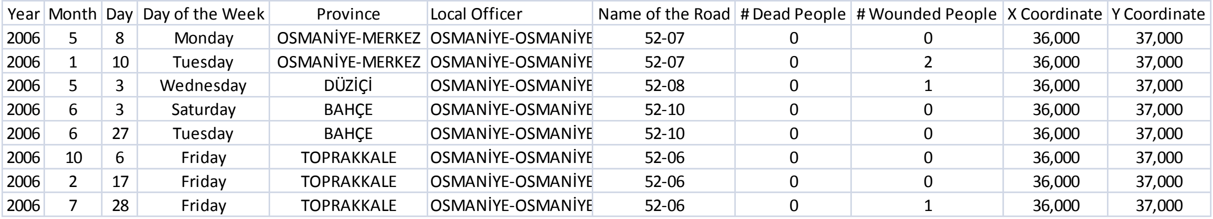
d. The coordinate information may be obtained before GPS devices get ready (GPS devices should receive information from at least three satellites for an accurate measure.)
e. The borders of a town or city may be changed or may have been defined incorrectly in the system's database.
f. Users may enter the same coordinates to the database systematically for any given points. The reasons may be numerous: users may not know the exact location; users may enter the same values to save time and it is easy; users may use dated reports to enter values. Such values are called as disguised missing values.

Disguised missing values are often insidious as they are hard to detect. Manual detection is hard and costly. To detect manually the densities of the hotspots can be identified and then can be analyzed further in detail whether they are legitimate or not. Here it is important to emphasize that disguised missing values are frequently entered values and they tend to be detected as hotspots in the analysis. Also one can look into the inconsistencies in the system. For example, the same coordinate may be entered so as to correspond to different towns or road names. Disguised missing data can appear in two different ways:

a. The coordinates may not be on the road and are too far away from the valid range (outliers).
b. Coordinates may appear as valid values i.e. on the road and within the valid range (town boundaries). This type of data is called as inliers.

We noticed that disguised missing values are entered in two different ways in the database:

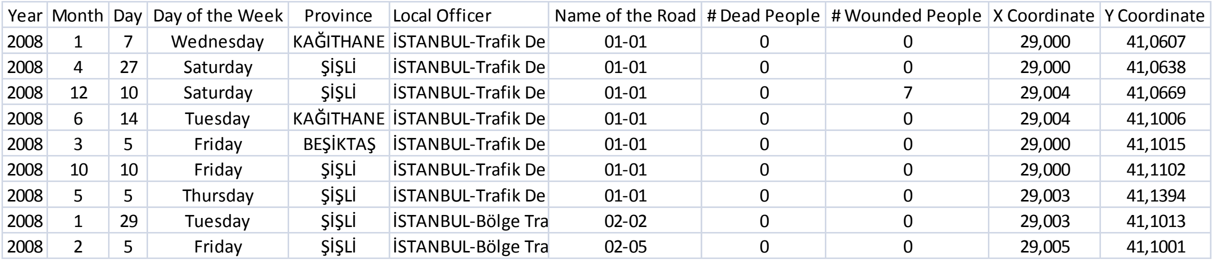
a. In the first case, the same coordinates are entered systematically to the system. For example, latitude 36 and longitude 37 are always entered instead of indicating the data entry as missing or unknown in city of Osmaniye in Turkey in year 2006 and 2007. Probably the users provided this well-known historical data in the cases where they cannot obtain the exact coordinates or they do not want to spend time to obtain them. You can see some part of the dataset in Table 2.



Year	Month	Day	Day of the Week	Province	Local Officer	Name of the Road	# Dead People	# Wounded People	X Coordinate	Y Coordinate
2006	5	8	Monday	OSMANİYE-MERKEZ	OSMANİYE-OSMANİYE	52-07	0	0	36,000	37,000
2006	1	10	Tuesday	OSMANİYE-MERKEZ	OSMANİYE-OSMANİYE	52-07	0	2	36,000	37,000
2006	5	3	Wednesday	DÜZİÇİ	OSMANİYE-OSMANİYE	52-08	0	1	36,000	37,000
2006	6	3	Saturday	BAHÇE	OSMANİYE-OSMANİYE	52-10	0	0	36,000	37,000
2006	6	27	Tuesday	BAHÇE	OSMANİYE-OSMANİYE	52-10	0	0	36,000	37,000
2006	10	6	Friday	TOPRAKKALE	OSMANİYE-OSMANİYE	52-06	0	0	36,000	37,000
2006	2	17	Friday	TOPRAKKALE	OSMANİYE-OSMANİYE	52-06	0	0	36,000	37,000
2006	7	28	Friday	TOPRAKKALE	OSMANİYE-OSMANİYE	52-06	0	1	36,000	37,000

Table 2: Highway Accidents Database show the accidents information for provinces in city of Osmaniye

b. Later case occurs due to GPS, manual entry or system problems. Users may enter spurious coordinate values with minor changes in the decimals. For example, in our accident data set, the points where Y coordinates vary between 41.050 and 41.102 and X coordinates vary between 29.000 and 29.005 are detected as disguised missing values. Even if it is not a particular point, these points which fall into a small area are still called disguised missing values. You can see the example in Figure 3.



Year	Month	Day	Day of the Week	Province	Local Officer	Name of the Road	# Dead People	# Wounded People	X Coordinate	Y Coordinate
2008	1	7	Wednesday	KAĞITHANE	İSTANBUL-Trafik De	01-01	0	0	29,000	41,0607
2008	4	27	Saturday	ŞİŞLİ	İSTANBUL-Trafik De	01-01	0	0	29,000	41,0638
2008	12	10	Saturday	ŞİŞLİ	İSTANBUL-Trafik De	01-01	0	7	29,004	41,0669
2008	6	14	Tuesday	KAĞITHANE	İSTANBUL-Trafik De	01-01	0	0	29,004	41,1006
2008	3	5	Friday	BEŞİKTAŞ	İSTANBUL-Trafik De	01-01	0	0	29,000	41,1015
2008	10	10	Friday	ŞİŞLİ	İSTANBUL-Trafik De	01-01	0	0	29,000	41,1102
2008	5	5	Thursday	ŞİŞLİ	İSTANBUL-Trafik De	01-01	0	0	29,003	41,1394
2008	1	29	Tuesday	ŞİŞLİ	İSTANBUL-Bölge Tra	02-02	0	0	29,003	41,1013
2008	2	5	Friday	ŞİŞLİ	İSTANBUL-Bölge Tra	02-05	0	0	29,005	41,1001

Table 3: Highway Accidents Database show the accidents information for provinces in city of İstanbul

Here, it is important to note that disguise values do not have to appear as unique values (as can be seen in Table 3) in spatial domain as they do in survey datasets. In survey datasets, values of an attribute can take a small number of values and one or two values are used as disguised missing values. However in spatial datasets, values of coordinates have much larger domain ranges and it is not plausible to work on only exact points as done in the first case.

Such values particularly when they are in large quantities may affect the outcome of the black spot analyses severely. A point in the area may be marked as black spot although it is not true.

The major inconsistencies regarding the coordinates that have been encountered in the dataset are as follows:

1. The coordinates are not on the road.
2. They may be entered as they belong to different towns in the same city.
3. They may be specified on different roads. For example, the coordinates shown in both Table 2 and Table 3 are entered as they belong to different roads, which is not possible.

Manual checking of all these inconsistencies require significant effort as each tuple in the database should be checked against any inconsistencies as well as on the map whether it is on the road or not.

4.1 Method

We have utilized the method proposed by Hua Ming et.al. (2007). However, we have made some adjustments on the method to improve the results and increase the performance. In the first section, the method will be explained and in the second section, we will describe the improvements we have made:

4.1.1. The Embedded Unbiased Sample (EUS) Heuristic

This approach is based on the heuristic that only a small number of values are frequently used as disguised missing values (one or two in an attribute) in real world data and these values are randomly distributed in the database.

If the GPS data is provided accurately to the system, we expect to see strong correlations between the coordinates and the town, coordinates and the name of the traffic inspectorate, coordinates and the road names, coordinates and the city. If a point is used as a disguise value such as [29;41.05] in city of İstanbul, then it will appear with different attributes values in the database i.e. the same coordinates with different towns, cities, and road names (see Table 3).

The tuples with the coordinates [29;41.05] will have both correctly and incorrectly recorded data.

$$\tilde{T}_{coordinates=[29.000-29.005;41.050-41.102]} = \tilde{R}_{coordinates=[29.000-29.005;41.050-41.102]} + S_{coordinates=[29.000-29.005;41.050-41.102]}$$

$\tilde{T}$is the recorded table, S is the table including disguised missing values and $\tilde{R}$ is the table with the true values.

EUS was defined as follows (Hua & Pei, 2007):
If v is a frequently used disguise value on attribute A, then $T_{A=v}$ contains a large subset $S_v \subseteq \widetilde{T}_{A=v}$ such that S_v is an ***unbiased sample*** *of $\tilde{T}$ except for attribute A where $\widetilde{T}_{A=v}$ = { $\tilde{t}\,\epsilon\check{\tilde{T}}$| $\tilde{t}$.A=v}.*

In order to clarify the example, let’s assume that we select tuples having towns as “Sarıyer” in city of İstanbul, then our subset will contain tuples associated with this specific region. We will observe *correlation* among some of the attributes i.e. only specific road names, name of the traffic inspectorate that records these values in the system will appear with “Sarıyer” in the database. Hence, this chosen subset will be a biased sample of the whole set. On the other hand, the subset including the coordinates [29;41.05] will be an unbiased sample if they appear randomly with other attribute values in the database. If they are in sufficient amount, they will be seen with every value of each attribute in the dataset.

If we select the tuples comprising v for the attribute A, it will contain both accurate tuples $\tilde{R}$ and contaminated tuples S_v. As a result, we aim to find the subset S_v from $\tilde{T}$ which is unbiased sample of the whole set. In other words, S_v will be the unbiased sample of $\tilde{T}$ except for the attribute A.

In order to find disguise values on an attribute A, it is required to find small number of attribute values whose projected databases contain a large subset as an unbiased sample of the whole table. Such attribute values are suspects of frequently used disguise values. The approach is straight forward; the larger the S_v the more suspicious for being a disguise value the v is. So we need to find **maximal unbiased subset *M_v***, maximal embedded unbiased ample, or MEUS for short.

While analyzing the subsets, two measures of M_v must be considered; size and quality. Quality can be defined as how well the subset resembles the distribution of the whole dataset. The values with large MEUS should be reported as the suspicious frequent disguise values. For the attributes that we suspect in the database, M_v will be calculated and the one with the largest value will be reported as disguised missing value.

The framework and details of the EUS heuristics can be found in (Hua & Pei, 2007)

4.1.2. Finding Disguised Missing Values in Spatial Datasets

EUS Heuristics have been successfully attempted on the health data sets. However, the method itself is not very appropriate for spatial datasets for the following reasons: Spatial datasets often include several different coordinate information. It is often costly to calculate the *correlation between each attribute values*. To gain from performance, we proposed to use "Chi Square Two Sample Test" instead of using correlation based scores.

Chi Square Two Sample Test checks whether two data samples come from the same distribution without specifying what that common distribution is. The chi-square two sample test is based on binned data. Binning for both data sets should be the same. The basic idea behind the chi-square two sample test is that the observed number of points in each bin (this is scaled for unequal sample sized) should be similar if the two data samples come from common distributions.

More formally, the chi-square two sample test statistic can be defined as follows.
H_0: The two samples come from a common distribution.
H_a: The two samples do not come from a common distribution.
Test Statistic: For the chi-square two sample tests, the data is divided into *k* bins and the test statistic is defined as:

$$x^2 = \sum_{i=1}^{k} \left(\frac{(K_1 R_i - K_2 S_i)^2}{R_i + S_i} \right)$$

where *k* is the number of categories (or bins), R_i is the observed frequency of bin *i* for the first sample, and S_i is the observed frequency of bin *i* for the second. K_1 and K_2 are scaling constants that are used to adjust for unequal sample sizes. Specifically,

$$\mathrm{K}_1 = \sqrt{\frac{\sum_{i=1}^{k} S_i}{\sum_{i=1}^{k} R_i}}$$

$$K_2 = \sqrt{\frac{\sum_{i=1}^{k} R_i}{\sum_{i=1}^{k} S_i}}$$

Our aim is to measure the distribution similarity between attribute couples of the dataset and the projected subset. In order to achieve this, we decided to represent the dataset as a means of *value couples* they include. For example we have generated new class labels for all the *value pairs* in attributes A_1 and A_2 that appear together. If A_1 and A_2 have *p* and *k* number of categories respectively, we have generated *pxk* new class labels, then created a new attribute column and finally put the new class labels that the value couples belong to. We repeated this data transformation for every attribute couple and for a dataset of *n* number of attributes we generated a new dataset *D* containing $\binom{n}{2}$ attributes. As we have generated the dataset which is appropriate for Chi Square Two Sample Test, we have measured the distribution similarity between class labels of the modified dataset and its projected subsets. The value which has the largest projected subset that has a similar distribution with the dataset

(Maximal Embedded Unbiased Sample, MEUS) is assigned as the most suspicious disguised missing value.

We have also taken into account the following constraints:

a. The uncorrelated or random attribute dimensions may affect the quality of the EUS heuristics based results negatively (Belen, 2009). One should select the related attributes with the help of an expert or using attribute selection algorithms. This also affects the performance of the method.

b. Binning coordinates is also important. It is necessary for Chi-Square Two Sample Test. We have also noticed that due to system, manual entry or GPS device problems, there are coordinates that are very close to each other but are disguised missing values. This is contrary to the belief that disguised missing value can often be a single value frequently found in a data set (Pearson, 2006).

Due to space limitation, we will not go into the details of these methods in this paper.

4.2 Experiment

4.2.1. Experiment Details and Data Set

Our data set comprises 22 attributes. Date', 'Time', 'State', 'Town', 'Number of people died', 'Number of people injured', 'X coordinate' and 'Y coordinate' are some of these attributes. Since the values of the coordinates were reported as unreliable by the experts, the disguised values were investigated on 'X coordinate' and 'Y coordinate' which correspond to the latitude and longitude coordinates respectively of the point on the highway where the accident happened.

There are 2 main problems: recorded coordinates are not on a highway and there are inconsistencies between the attribute values. In the former case, recorded coordinates fall into terrain or city centers rather than highways. In the latter case, coordinates are specified on the highways but there are inconsistencies with other attribute fields.

As a point is represented by two coordinates, we should take into consideration both values together instead of analyzing individual attributes while detecting disguised missing values. So we generated new point labels for each *x* and *y* couples.

For a dataset with *n* number of values for *x* coordinate and *m* number of values for *y* coordinate, we obtained *mxn* point labels. Handling hundreds of points is not sensible in spatial data sets. Although many accidents may happen on the same place, coordinates might not be recorded exactly the same but recorded with minor differences. Therefore it is required to work on disguised values of *areas* not the *points*. So what we need to know is these areas that are frequently used as disguised values.

We can form regions based on the coordinates. It is better to create these regions with a domain expert if it is possible for more accurate analysis. Here it is important to note that we did not include any domain knowledge like road segments or valid X or Y coordinate ranges that define the boundaries of a city.

As a coordinate has 2 to 4 decimals, we have decided to create regions by using a binning range (e.g. 10, 100). To be on the safe side, we rounded the values by 10, 100, and 1000. We also have run the algorithm without rounding the values as well.

Depending on the characteristics of the dataset, different binning ratios may lead to more appropriate results. In our case rounding by 100 has worked satisfactorily for each dataset.

Since the approach is based on the correlation between attribute couples, we should attach importance to which attributes to cover. Redundant attributes can be eliminated by using feature selection methods such as minimum-redundancy maximum-relevancy algorithm. The goal of the algorithm is to select a feature subset that best characterizes the statistical property of a target classification variable, subject to the constraint that these features are mutually as dissimilar to each other as possible, but marginally as similar to the classification variable as possible (Peng, Long, & Ding, 2005). If we select the classification variable as A_v, the attribute comprising the candidate disguised value, we can find out the attributes that are dissimilar to A_v and remove them from the calculations. As a consequence, we will eradicate independent attributes which cause biased results. The method has returned the following features as related: “Official Report Number”, “Name of the Officer”, “Road ID”, “Town” and “Day of the Week”.

Selecting the feature subset also increases the computational efficiency which has a vital effect while working with large datasets. Since we are dealing with value sets of attributes, dimension reduction decreases the time required to run the method exponentially.

In our analysis, we worked on the accidents occurred in the cities of Adana, Gaziantep, Hatay, Mersin and Osmaniye in 2006, 2007 and 2008.

Year	Adana	Gaziantep	Hatay	Mersin	Osmaniye	Istanbul
2006	141	82	32	152	108	
2007	147	119	25	160	92	
2008	118	137	38	170	95	1379

Table 4: Accidents occurred between 2006-2008

4.2.2. Results

We have consulted to the domain expert to evaluate the results. Our analysis for accidents occurred in Adana in 2006 and 2007 did not come up with a frequent suspicious disguised missing value. However, we have detected an area as a candidate disguised missing area in 2008. There are 11 accident points recorded within the area where X coordinate ranges between 35.746 to 35.754 and Y coordinate ranges between 36.976 to 36.984 (X coordinate is rounded to 35.75 and Y coordinate is rounded to 36.98). These points are on the highways segments. However they contradict with other attributes. For example, coordinate values have been recorded to fall into 3 different roads. Two of these road segments (road names 52-03 and 52-04) are successive but one of them (21-05) is totally irrelevant (see Figure 1). We have discussed our findings with the domain expert and he has confirmed that these values exhibit inconsistencies suggesting that this area should be marked as disguise value.

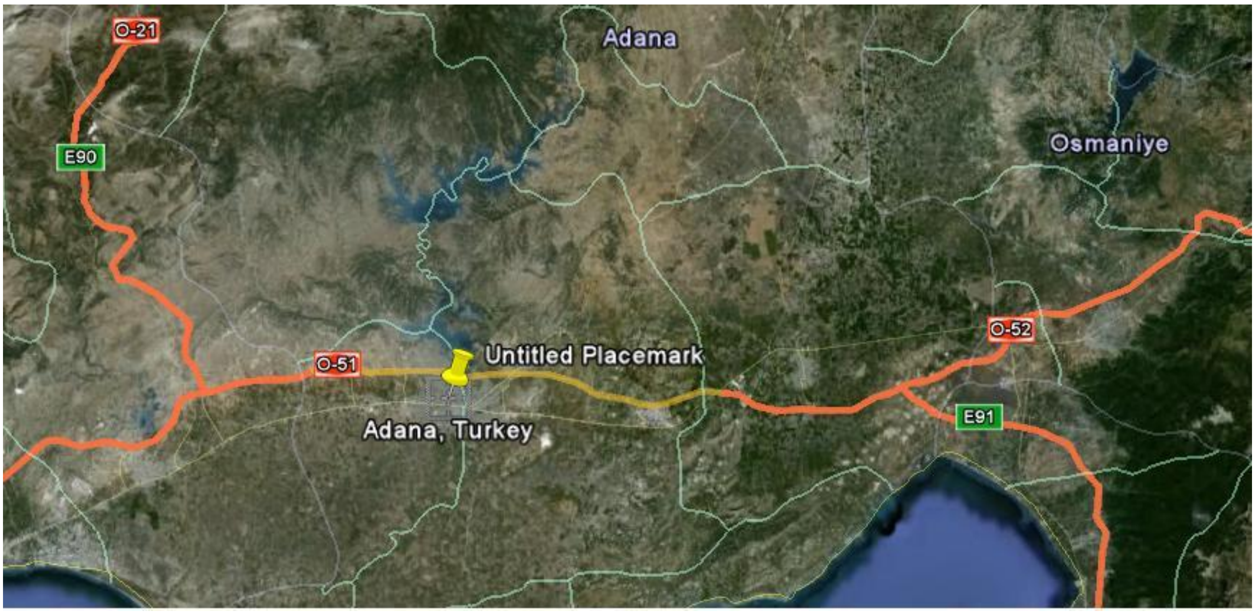


Figure 1: 11 accident black spots between (35.7507; 36.9881) and (35.7509; 36.9899) were specified on O-21 and O-52 roads in the data set of city of Adana in 2008. The black spots are denoted as a yellow placemark on the map. Due to the resolution, it appears to be one black spot. But there are 11 different points which are very close to each other. Table 2 shows some of these tuples. The area includes disguise values. The problem here lies in the road names. The same black spots are indicated as they belong to different roads that is not possible. **@Courtesy of Google Earth.**

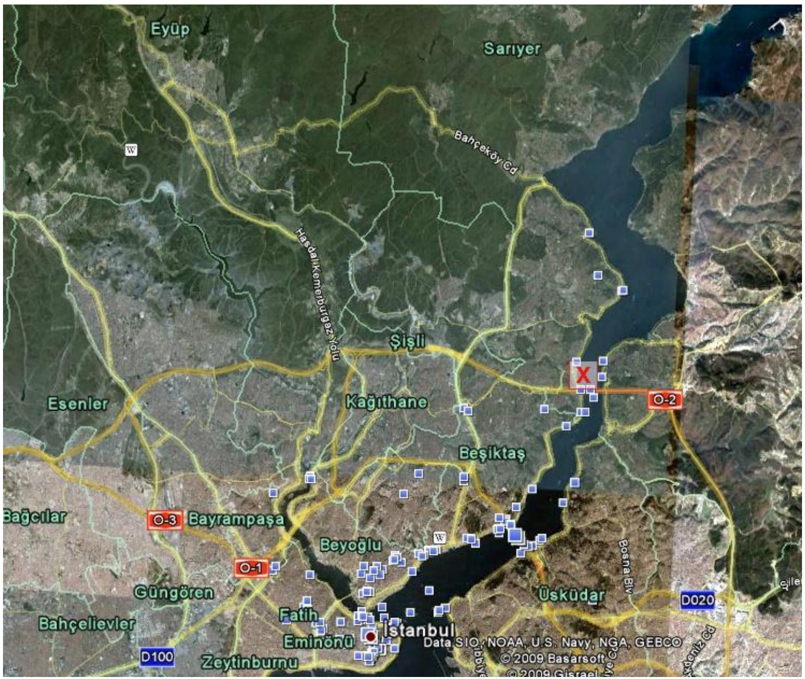


Figure 2: The figure shows a part of the traffic accidents recorded in İstanbul in 2008. A potential disguise value is marked as red cross on the map. This point is entered as it is within the boundaries of Kağıthane, Şişli, Beşiktaş and Sarıyer. However, the road name is specified as 01-01 in the database (the first part corresponds to the road number and the second part implies at which kilometer of the road the accident happened such as here the accident happened in the first kilometer of the road 01). The road 01-01 cannot belong to four different towns. Table 3 shows some of these tuples. **@Courtesy of Google Earth.**

Figure 2 shows the road accidents in İstanbul. We have noticed 27 points were entered in the similar whereabouts i.e. latitude varies between [29-29.005] and longitude varies between

[41.05-41.102] in the database. Although these coordinates are only within the boundaries of only one district, four different district names are entered.

Our algorithm returned several areas as the suspicious disguise missing areas for İstanbul. Some of them are not the points on the highways and are consequently approved by the domain expert. We have detected 5 suspicious areas. Three of them (31 tuples in total) are expressed as the areas outside the highways by the domain expert. In two of them, we have observed many inconsistencies. In one of the areas whose X coordinates range between 29 to 29.0046 and Y coordinates range between 41.0597 to 41.1001, 35 accident points are recorded on 2 different highways (01-01 and 02-05) which is impossible (see Figure 2). In addition, the coordinates are specified as belonging to four different districts. Such an abnormality shows that points within that range are not the actual accident points.

Our analysis did not come up with a frequent suspicious disguised missing value for Gaziantep, Mersin and Hatay in any year. When we have consulted to domain expert about this finding, he has confirmed the results and explained that these datasets are reliable.

The most remarkable results are gathered in the state Osmaniye in 2006. In Osmaniye, 66 accidents out of 108 have been recorded to occur at the coordinate (36.00, 37.00) which is apparently suspicious. This value is recorded on 5 different highways and 4 towns. In 2007, the same coordinates are observed in 26 accidents out of 92 accidents.

4.3 Conclusion

Highway accident data sets provide valuable information regarding the detection of the black spots on the road and thus aid officers to take timely precautions to prevent accidents from happening in the future. However, these results will be trustable if we only have high quality data. In this paper, we have investigated a data quality problem which is the identificaton of the disguised missing data in the recorded accident data sets of six cities in Turkey. We have found that in the data sets of three cities, there are disguised missing data values.

Since we do not follow up any domain constraint, we do not claim that we can find any point which is *not* on the highway. However, we can handle a more insidious case in which we can find the disguised missing values that can be inlier or outlier.

Here, we want to emphasize that if an area is frequently selected as disguised missing value, it is very likely to be detected as a hotspot. Therefore it is essential to uncover disguised missing values before identifying the hotspots. Otherwise, decisions made based on these counterfeit hotspots will be worthless and perversive. While detecting disguised missing values, inconsistencies in the tuples are also uncovered which clearly show the problems.

4.4 Discussion and Future Work

While it is impossible to prevent disguised missing values in a dataset, it is possible to prevent users to enter inconsistent values to some extent. Users may tend to enter fake values for various reasons when they are forced to provide a value. Consequently, in particular for such critical and mandatory fields, it may be useful to restrict users to enter a value only within a valid range. For example, when a user enters the town in which the accident occurred, it will be helpful to narrow down the roads to the ones that are in that town and asking user to select one of the given roads. Such guidance will both help the user and motivate him/her to enter actual values. However these approaches do not ensure to prevent all the disguised missing values from happening but they help to eliminate some of them. As we mentioned earlier, we do not include any domain knowledge. For a more accurate result, it may be appropriate to eliminate the points that are not on the highway first and analyzing the disguised missing values later. At that point, algorithm will only return the inliers that are impossible to detect manually.

REFERENCES

Adverse Event Reporting System (AERS). (2009, 1). Retrieved 1 2009, from U.S. Food and Drug Administration: http://www.fda.gov/cder/aers/default.htm

Belen, R. (2009). *Detecting Disguise Missing Data.* METU. Ankara, Turkey: Unpublished MSc Thesis.

Berti-Equille, L. (2007). Data Quality Awareness: A case study for cost-optimal association rule mining. *Knowledge and Information Systems , 11* (2), 191-215.

Canadian Council of Motor Transport Administrators. (2009). *Road Safety Vision 2010*. Retrieved 10 30, 2009, from http://www.ccmta.ca/english/committees/rsrp/rsv/rsv.cfm

Carmel, L. (2008, 5). *Multivariate Analysis Toolbox for Matlab®* . Retrieved 1 2009, from http://www.ncbi.nlm.nih.gov/CBBresearch/Fellows/Carmel/software/MVA/mva.html

Chouinard, A., & Lecuyer, J.-F. (2009). Achieving quality of Canadian crash data. *4th IRTAD Conference on Road safety data: collection and analysis for target setting and monitoring performances and progress.* Seoul, Korea: International Transport Forum.

Cinzia Cappiello, C. F. (2004). Data Quality Assessment from the User's Perspective. *Proceedings of the 2004 international workshop on Information quality in information systems* (pp. 68-73). ACM New York, NY, USA.

Dasu T., J. T. (2003). *Exploratory Data Mining and Data Cleaning.* New York: Wiley-Interscience.

Databank. (2006, 1). Retrieved 1 2009, from Turkish Statistical Institute: http://www.tuik.gov.tr/jsp/duyuru/upload/vt/vt.htm

Elsig, K. (2009, 5). *Road crash and road crash injury data for setting and monitoring targets.* Retrieved 10 30, 2009, from UNECE Seminar on Improving Global Road Safety: http://www.unece.org/trans/roadsafe/unda/Minsk_Pres18_Elsig.pdf

Heckert, A. (2006). *Chi square two sample.* Retrieved 1 2009, from http://www.itl.nist.gov: http://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/chi2samp.htm

Heiko Müller, J.-C. F. (2003). Problems, Methods, and Challenges in Data Cleansing. Berlin: HUB-IB-164.

Hipp, J., Güntzer, U., & Grimmer, U. (2001). Data Quality Mining: Making a Virtue of Necessity. *Proceedings of the 6th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, (pp. 52-57). Tübingen.

Hua, M., & Pei, J. (2007). Cleaning Disguised Missing Data: A Heuristic Approach. California: KDD' 07.

Hulse, J. D., Khoshgoftaar, T. M., & Huang, H. (2007). The Pairwise Attribute Noise Detection Algorithm. *Knowledge and Information Systems , 11* (2), 171-190.

Inaccurate Customer Data Results In Lost Revenue. (2005, 06 09). Retrieved 01 2009, from Experian QAS: http://www.qas.com/display-news.htm?id=4731

International Traffic Safety Data and Analysis Group. (2009). Seoul Statement. *4th IRTAD Conference Road Safety Data: Collection and Analysis for Target Setting and Monitoring Performance and Progress.* Seoul.

Jochen Hipp, U. G. (2001). Data Quality Mining: Making a Virtue Neccessity.

Leo L. Pipino, Y. W. (2002). *Data Quality Assessment* (Vol. 45). New York, NY, USA: ACM.

Leur, P. d., & Sayed, T. (2002). Development of a Road Safety Risk Index. *Transportation Research Record: Journal of the Transportation Research Board , 1784* (2002), 33-42.

Markets, I. r. (2005). *How well do you know your customers?* Experian Press.

Maydanchik, A. (2007). *Data Quality for Practitioners: Data Quality Assessment.* Technics Publications, LLC.

Olson, J. E. (2003). *Data Quality: The Accuracy Dimesion.* San Francisco: Morgan Kaufmann.

Pearson, R. (2006). The Problem of Disguised Missing Data. *ACM SIGKDD Explorations Newsletter , 8* (1), 83-92.

Peden, M., & Toroyan, T. (2005). Counting Road Traffic Deaths and Injuries: Poor Data Should Not Detract From Doing Something! *Annals of Emergency Medicine , 46* (2).

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence , 27* (8), 1226-1238.

Pima Indians Diabetes Data Set . (n.d.). Retrieved 1 2009, from UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

Redman, T. C. (1997). *Data Quality for the Information Age.* Artech House Publishers,Norwood, MA, USA.

Richard D. De Veaux, D. J. (2005). How to Lie with Bad Data. *20.*

Salifu, M., & Ackaah, W. (2009). Under-reporting of road traffic accident data in Ghana. *4th IRTAD Conference on Road safety data: collection and analysis for target setting and monitoring performances and progress.* Seoul, Korea: International Transport Forum.

State Univ. of New York at Albany. (1998). Data Quality Tools for Data Warehousing- A Small Sample Survey. Albany: Defense Technical Information Center.

Stuart Madnick, H. Z. (2005). Improving Data Quality Through Effective Use of Data Semantics. *Data and Knowledge Engineering , 59* (2), 460-475.

Transportation Research Board National Research Council. (1991). *Accident Data Quality: A synthesis of highway practice.* Washington, D.C.: NCHRP Synthesis 192.