Metodološki zvezki, Vol. 11, No. 2, 2014, 107-120

A Distance Based Measure of Data Quality

Pavol Kráľ[1], Lukáš Sobíšek[2], Mária Stachová[3]

Abstract

Data quality can be seen as a very important factor for the validity of information extracted from data sets using statistical or data mining procedures. In the paper we propose a description of data quality allowing us to characterize data quality of the whole data set, as well as data quality of particular variables and individual cases. On the basis of the proposed description, we define a distance based measure of data quality for individual cases as a distance of the cases from the ideal one. Such a measure can be used as additional information for preparation of a training data set, fitting models, decision making based on results of analyses etc. It can be utilized in different ways ranging from a simple weighting function to belief functions.

1 Introduction

According to Cox (Cox 1972) "issues of data quality and relevance, while underemphasized in the theoretical statistical and econometric literature, are certainly of great concern in much statistical work". Nevertheless, data quality issues are mostly discussed in connection to data collection, data storage and data extraction and preparation processes, not statistical and data mining procedures themselves. In the presented paper we focus on data quality as a possible input for further data analysis and/or decision making based on results of this analysis. The main goal is to propose a simple and easily applicable measure for data quality. In our opinion, such a measure should aggregate various aspects of data quality, for example completeness, uncertainty, imprecision etc. (Berti-Equille 2007, Parsons 1996). Assuming that each aspect of data quality for a particular data entry can be assessed by a single number from the unit interval, data quality of a particular variable can be expressed by a corresponding n-tuple of mappings where each mapping maps values of a variable recorded in a data set into the unit interval. Data quality of a particular case then aggregates data quality of all corresponding variables in the form of a family of n-tuples. If we use the above mentioned data quality description, it allows us to represent data quality of a case as a distance from the ideal case, i.e. the case without any data imperfection. In the rest of our paper we call such a distance Data Quality Index and denote it DQI. The DQI can be used as prior information for further modelling (classification,

[1] Department of Quantitative Methods and Information Systems, Faculty of Economics, Matej Bel University, Tajovského 10, 975 90 Banská Bystrica; pavol.kral@umb.sk

[2] University of Economics, Prague W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic; lukas.sobisek@vse.cz

[3] Department of Quantitative Methods and Information Systems, Faculty of Economics, Matej Bel University, Tajovského 10, 975 90 Banská Bystrica; maria.stachova@umb.sk

clustering etc.), e.g. in the form of weights for particular cases. Instead of using DQI as a direct input for our analyses, we can use it as a source for measuring data quality of the whole data set. Data quality (reliability, validity) of the whole data set can be then used as a supplement to decision making based on results of statistical analysis.

The paper is organized as follows. In Section 2 we review data quality issues discussed in literature. Section 3 forms the main part of the paper: first we propose data quality description of individual variables and statistical units, then, on the basis of this representation, we construct DQI, a simple real valued measure of data quality for statistical units. Finally, in Section 4 we apply the proposed distance based data quality measure to a real data set.

2 Data quality

Data quality is a term with very broad meaning. In (Berti-Equille 2007) the author presents the following main data quality issues: duplicate and redundant data, imperfect data with low accuracy, missing values and incomplete databases and stale, i.e. non-fresh data. It is obvious that importance of these particular aspects of data quality depends on the problem we are trying to solve, what are our goals, what methods we intend to use, whether a particular data issue can be solved etc. For example, duplicate and redundant data can be effectively handled by fusion or deletion of records in the process of data extraction from a warehouse and many authors described how to deal with missing values and incomplete data in the past (see Imielinski and Lipski 1984, Grahne 2002 and Naumann, Leser, Freytag 1999). Contrary, in the case of two remaining families of data quality issues, data freshness and data accuracy, it is quite impossible to deal with them prior to the assumed analysis. Therefore we focus on them in the rest of our paper.

2.1 Data Freshness

Segev and Fang in (Segev, Fang 1990); Theodoratos and Bouzeghoub in (Theodoratos, Bouzeghoub 1999) use the traditional freshness definition called currency. It takes into account the difference between Query Time[1] and Extraction Time[2]. Another notion of freshness, called timeliness, describes the ageing of data. It describes how often data changes, it means it takes into account the difference between Query Time and Last Update Time[3] (Naumann, Freytag, Leser 2004).

The freshness factors and their corresponding metrics, summarized in (Peralta 2006), are listed in Table 1.

The relevance of data freshness factors and metrics from the point of view of statistical analysis depends on goals of analysis. For example, it is more relevant for frequent basic reporting than for supervised learning.

[1] Query Time is the instant time, when users retrieve data.
[2] Extraction Time refers to the starting time, when extracted data is used.
[3] Last Update Time corresponds to the time, when data was last updated.

Table 1: Summary of freshness factors and metrics.

Factor	Metric	Description
Currency	Currency	The time elapsed since data was extracted from the source (the difference between the delivery time and extraction time).
	Obsolescence	The number of updates operations to a source since the extraction time.
	Freshness ratio	The percentage of tuples in the view that are up-to-date.
Timeliness	Timeliness	The time elapsed from the last update to a source (the difference between the delivery time and last update time).

2.2 Data accuracy

Data accuracy plays a key role in data quality studies. Data with low accuracy can be defined as imperfect data. This is a very broad term further characterized by Parsons (Parsons 1996). Parsons compiles earlier works of Bonnissone and Tong (Bonnissone, Tong 1985), Bosc and Prade (Bosc, Prade 1993), and splits imperfect data into five separate parts, namely incomplete information, uncertainty, imprecision, vagueness and inconsistency. Moreover, Parsons specifies the above mentioned terms, describes their sources and offers solutions how to deal with these issues.
In our analysis we focus on uncertainty in data. Motro (Motro 1993) claims "Uncertainty permeates our understanding of the real world. The purpose of information systems is to model the real world. Hence information systems must be able to deal with uncertainty." If a system provides poor data to data users (analysts, researchers), they must incorporate uncertainty into their modelling strategies.

It is obvious that this factor cannot be easily exactly defined. It is strictly context dependent and has to be evaluated with respect to the analyzed problem. It can include expert information and intuitive approach based on users' (analysts) expectations and combine them with exact statistical techniques (e.g. clustering, classification, regression,...).

3 Data quality description and a distance based data quality measure

As it was mentioned in the previous chapter, data quality attributes are often context dependent. In our opinion, regardless the problem we are trying to solve, data quality can be viewed from the three different perspectives:

1. data quality of variables,
2. data quality of particular cases (statistical units),
3. data quality of a data set.

We can evaluate different data quality attributes (uncertainty, freshness, missingness etc.) from a local or global point of view. The global view means that we are able to decide whether the examined variable or statistical unit is appropriate for our analysis, e.g. we can remove variables and statistical units with high missingness or penalize variables with high uncertainty. The local view means that we are interested in data quality of a variable (a statistical unit) for a particular statistical unit (a particular variable), e.g. data entries for a particular case were made just a moment before a data set extraction, therefore freshness of this variable for that particular case is very good. The local view can be used to decide if a statistical unit would be used for our analysis unchanged, penalized or boosted. Obviously, if we aggregate local data quality of a variable for all available statistical units, we get global data quality of this variable. Analogously, if we aggregate local data quality of all variables for a statistical unit, we get global data quality of this statistical unit. On the other hand, we are often able to assess global data quality of a variable (a statistical unit) without aggregating its local data quality for statistical units (variables).

In the rest of our paper we assume for simplicity that we work with data sets already prepared for analysis, i.e. variables with the high number of missing values were already removed, duplicity in data entries was resolved etc. Moreover, we do not deal with variables with obvious 100 % data quality, i.e. variables without uncertainty, irrelevant freshness etc. Gender is an example of such a variable. It means that we focus primarily on the local view of data quality.

3.1 Data quality description

The basic element of our data quality description is formed by the definition of data quality of a variable with respect to a chosen data quality attribute (freshness, uncertainty, etc.) or a set of attributes, and a particular data set.

Definition 1. Let X denotes an observed variable, A denotes an attribute of data quality and $\mathcal{C}$ denotes a set of statistical units. Then the data quality of X with respect to A and $\mathcal{C}$ is a mapping $D_{X,A,\mathcal{C}} : \mathrm{ran}(X) \times \mathcal{C} \to [0,1]$, where $\mathrm{ran}(X)$ denotes the range of the variable X. If $\mathrm{ran}(D_{X,A,\mathcal{C}}) = \{1\}$, the variable X has 100 % data quality with respect to the attribute A and the set $\mathcal{C}$. If $\mathrm{range}(D_{X,A,\mathcal{C}}) = \{0\}$, the variable X has 0 % data quality with respect to the attribute A and the set of statistical units $\mathcal{C}$.

Clearly, even if the variable X takes the same value for two statistical units $c, c' \in \mathcal{C}$, the mapping $D_{X,A,\mathcal{C}}$ can take completely different values. Definition 1 can be generalized to a set of attributes in the following way.

Definition 2. Let X be an observed variable, $\mathcal{A} = \{A_1, A_2, \ldots, A_p\}$ be a set of data quality attributes and $\mathcal{C}$ be a set of statistical units. Then the data quality of X with respect to $\mathcal{A}$ and $\mathcal{C}$ is a p-tuple

$$\left(D_{X,A_1,\mathcal{C}}, D_{X,A_2,\mathcal{C}}, \ldots, D_{X,A_p,\mathcal{C}}\right), \tag{3.1}$$

where $D_{X,A_i,\mathcal{C}}$ denotes the data quality of X with respect to the attribute A_i and the set of statistical units $\mathcal{C}$.

Using the data description of variables from Definition 1 and 2 we can characterize data quality of a particular case (a statistical unit) $c \in \mathcal{C}$ with respect to a variable X and a set of attributes $\mathcal{A}$.

Definition 3. Let $\mathcal{C}$ be a set of statistical units, X be a variable and $\mathcal{A} = \{A_1, A_2, \ldots, A_p\}$ be a set of data quality attributes. Then the data quality of a statistical unit $c \in \mathcal{C}$ with respect to X, $\mathcal{A}$ and $\mathcal{C}$ is a p-tuple defined as follows

$$\left(D_{X,A_1,\mathcal{C}}(x,c), D_{X,A_2,\mathcal{C}}(x,c), \ldots, D_{X,A_p,\mathcal{C}}(x,c)\right), \tag{3.2}$$

where x is a value of X measured on c.

For simplicity, we denote the p-tuple (3.2) by $D_{X,\mathcal{A},\mathcal{C}}$.

The previous definition can be straightforwardly extended to a set of variables as follows.

Definition 4. Let $\mathcal{C}$ be a set of statistical units, $\mathcal{X} = \{X_1, X_2, \ldots, X_m\}$ be a set of variables and $\mathcal{A} = \{A_1, A_2, \ldots, A_p\}$ be a set of data quality attributes. Then the data quality of a statistical unit $c \in \mathcal{C}$ with respect to $\mathcal{X}$, $\mathcal{A}$ and $\mathcal{C}$ is an m-tuple

$$(D_{X_1,\mathcal{A},\mathcal{C}}, D_{X_2,\mathcal{A},\mathcal{C}}, \ldots, D_{X_m,\mathcal{A},\mathcal{C}}). \tag{3.3}$$

In the following example we illustrate Definitions 1-4.

Example 1. Let us assume $\mathcal{C} = \{c_1, c_2, c_3\}$, $\mathcal{X} = \{X_1, X_2\}$ and $\mathcal{A} = \{A_1, A_2\}$. Moreover, let $X_1(c_1) = x_{11}$, $X_1(c_2) = x_{11}$, $X_1(c_3) = x_{13}$, $X_2(c_1) = X_2(c_2) = X_2(c_3) = x_{21}$. Then, according to Definition 1, the mappings

$$D_{X_1,A_1,\mathcal{C}} = \begin{cases} 0.5 & \text{for } (x_{11}, c_1), \\ 0.4 & \text{for } (x_{11}, c_2), \\ 0.7 & \text{for } (x_{13}, c_3), \\ 0 & \text{elsewhere,} \end{cases}, \quad D_{X_1,A_2,\mathcal{C}} = \begin{cases} 0.2 & \text{for } (x_{11}, c_1), \\ 0.8 & \text{for } (x_{11}, c_2), \\ 0.6 & \text{for } (x_{13}, c_3), \\ 0 & \text{elsewhere,} \end{cases}$$

$$D_{X_2,A_1,\mathcal{C}} = \begin{cases} 0.3 & \text{for } (x_{21}, c_1), \\ 0.5 & \text{for } (x_{21}, c_2), \\ 0.4 & \text{for } (x_{21}, c_3), \\ 0 & \text{elsewhere,} \end{cases}, \quad D_{X_2,A_2,\mathcal{C}} = \begin{cases} 0.1 & \text{for } (x_{21}, c_1), \\ 0.2 & \text{for } (x_{21}, c_2), \\ 0.9 & \text{for } (x_{21}, c_3), \\ 0 & \text{elsewhere,} \end{cases}$$

are examples of data quality of X_1 and X_2 with respect to A_1, C and A_2, C. Applying $D_{X_1,A_1,\mathcal{C}}$, $D_{X_1,A_2,\mathcal{C}}$, $D_{X_2,A_1,\mathcal{C}}$, $D_{X_2,A_2,\mathcal{C}}$ we get the following data quality of X_1 and X_2 with respect to $\mathcal{A}$ and $\mathcal{C}$ (see Definition 2):

$$(D_{X_1,A_1,\mathcal{C}}, D_{X_1,A_2,\mathcal{C}}) \text{ and } (D_{X_2,A_1,\mathcal{C}}, D_{X_2,A_2,\mathcal{C}}).$$

Then, following Definition 3, for data quality of the statistical unit c_1 it holds, with respect to X_1, $\mathcal{A}$ and $\mathcal{C}$,

$$(D_{X_1,A_1,\mathcal{C}}(x_{11}, c_1), D_{X_1,A_2,\mathcal{C}}(x_{11}, c_1)) = (0.5, 0.2)$$

and, with respect to X_2, $\mathcal{A}$ and $\mathcal{C}$,

$$(D_{X_2,A_1,\mathcal{C}}(x_{21}, c_1), D_{X_2,A_2,\mathcal{C}}(x_{21}, c_1)) = (0.3, 0.1)$$

Analogously, we get, with respect to X_1, $(0.4, 0.8)$ for c_2 and $(0.7, 0.6)$ for c_3. With respect to X_2, we get $(0.5, 0.2)$ for c_2 and $(0.4, 0.9)$ for c_3. Finally, data quality of c_1, c_2 and c_3 is the following, with respect to $\mathcal{X}$, $\mathcal{A}$ and $\mathcal{C}$,

$((0.5, 0.2), (0.3, 0.1))$, $((0.4, 0.8), (0.5, 0.2))$ and $((0.7, 0.6), (0.4, 0.9))$, respectively.

Data quality description of variables and statistical units can be a basis for data quality description of the whole data set. Let the dimension of the whole data set be $n \times m$. Then data quality of the whole data set can be characterized either as an n-tuple, where each element represents data quality of a particular case (statistical unit), or as an m-tuple, where each element represents data quality of a particular variable. The above mentioned data quality description is exhaustive, incorporates data quality of all variables and statistical units with respect to any set of attributes. Unfortunately, from the practical point of view our description is not easily applicable (large dimensions, complex interpretation etc.). Therefore we construct on its basis a simple data quality measure aggregating the complete description of data quality of each statistical unit or statistical variable into a single real number.

3.2 A distance based data quality measure – Data Quality Index

There are many possibilities how to use our data quality description as a basis for further analysis or as additional information which supplements results of our analysis. Because we do not assume that attributes of data quality are independent we restrict ourselves to the distance based data quality measure, DQI. It means that data quality of a statistical unit is defined as a distance of its m-tuple from an m-tuple describing the ideal statistical unit, i.e. the statistical unit without any data quality issues. In our paper the term distance coincides with the term metric, i.e. we require its non-negativity, identity of indiscernible, symmetry and triangle inequality.

Definition 5. Let $\mathcal{C}$ be a set of statistical units, let data quality of each statistical unit $c \in \mathcal{C}$ be described by (3.3), let d be a distance function. Then a mapping $\mathrm{DQI}\colon \mathcal{C} \to [0, 1]$ is defined as follows

$$\mathrm{DQI}(c) = d((D_{X_1,\mathcal{A},\mathcal{C}}, D_{X_2,\mathcal{A},\mathcal{C}}, \ldots, D_{X_m,\mathcal{A},\mathcal{C}}), \mathbf{1}), \tag{3.4}$$

where $\mathbf{1}$ denotes the m-tuple $\left(\underbrace{(1, 1, \ldots, 1)}_{p}, \ldots, \underbrace{(1, 1, \ldots, 1)}_{p}\right)$.

DQI can take values from the unit interval, where 0 means that a statistical unit c has not any data issues with respect to $\mathcal{A}$ and 1 means that a statistical unit c has 0 % data quality.

Data quality of the whole data set we can characterize as a sum or an appropriate measure of central tendency, e.g. mean or median, of all $\mathrm{DQI}(c)$, where $c \in \mathcal{C}$.

We have many possibilities how to choose an appropriate distance used in formula (3.4). It is similar to selecting an appropriate distance in the case of clustering. It is obvious that our data quality description of a statistical unit is mathematically equivalent to so called hesitant fuzzy sets (Zeshui, Meimei 2011), although its interpretation is completely different. Therefore, in the rest of our paper, we restrict ourselves to two distances similar to those used in the case of hesitant fuzzy sets, the normalized Hamming like distance

$$d_{NHD}(c,c') = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{p}\sum_{j=1}^{p}|D_{X_i,A_j,\mathcal{C}}(x_i,c_i) - D_{X_i,A_j,\mathcal{C}}(x'_i,c'_i)|\right], \tag{3.5}$$

and the normalized Euclidean like distance

$$d_{NED}(c,c') = \left(\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{p}\sum_{j=1}^{p}\left(D_{X_i,A_j,\mathcal{C}}(x_i,c_i) - D_{X_i,A_j,\mathcal{C}}(x'_i,c'_i)\right)^2\right]\right)^{\frac{1}{2}}, \tag{3.6}$$

where $c, c' \in \mathcal{C}$.

Remark Similarly, we can introduce DQI for variables as a mapping $\mathrm{DQI_v}\colon \mathcal{X} \to [0,1]$:

$$\mathrm{DQI}(X) = d((D_{X,\mathcal{A},c_1}, D_{X,\mathcal{A},c_2}, \ldots, D_{X,\mathcal{A},c_n}), \mathbf{1}),$$

where $\mathbf{1}$ denotes the n-tuple $\left(\underbrace{(1,1,\ldots,1)}_{p},\ldots,\underbrace{(1,1,\ldots,1)}_{p}\right)$, $D_{X,\mathcal{A},c_i}$ denotes a p - tuple $\left(D_{X,\mathcal{A}_1,\mathcal{C}}(X(c_i),c_i), D_{X,\mathcal{A}_2,\mathcal{C}}(X(c_i),c_i),\ldots,D_{X,\mathcal{A}_p,\mathcal{C}}(X(c_i),c_i)\right)$ and d denotes a distance function. Therefore we can apply our data quality measuring algorithm to statistical units as well as to statistical variables.

Our approach is roughly inspired by the TOPSIS method (Hwang, Yoon 1981) but our algorithm assumes the best alternative independently of existing statistical units and does not assume the worst alternative. Moreover, in the case of TOPSIS method we are interested in ranking of alternatives in order to choose the best alternative, in our approach we are interested in an absolute measure of data quality of a particular statistical unit allowing us to decide whether and how this statistical unit can be used in our further analyses. On the other hand, similarly to TOPSIS, our method allows a trade-off between data quality of attributes, where one attribute can be compensated by another one. The level of compensation depends on the number of attributes and variables.

4 Application of data quality analysis to insurance data

The application of data quality analysis depends strongly on the studied research problem. Nevertheless, we can pose some recommendations how to incorporate data quality analysis into a data analysis process. In order to illustrate such possible inclusion, we present here partial data quality analysis in the context of statistical analysis we performed on a real data set. The data set comes from a Czech insurance company and consists of 677,284

real customer contracts (units, i.e. rows) with 9 characteristics (variables, i.e. columns). One of these characteristics (the dependent variable) represents classification of the contract and other variables are described in Table 2. 261,402 cases belong to the customers with a lapsed insurance policy and 415,882 cases to the customers with a policy in force. Although we fully support the idea of reproducible research, the insurance company did not give us permission to share data in any form due to its confidentiality policy.

Table 2: Description of used variables and their notation.

Variable	Type
type of product	dichotomous
payment frequency (within one year)	categorical with 5 levels
region	categorical with 14 levels
gender	dichotomous
the age of policyholder at the time of conclusion of the contract (in years)	numeric
number of policyholder migrations	categorical with 11 levels
freshness (in years)	numeric
policy duration (in years)	numeric

Variables listed in Table 2 can be classified into 2 types: policyholder's characteristics (age, gender, region and number of migrations) and contract's characteristics (product type, payment frequency, freshness and policy duration). The first step of our analysis consists of elements of exploratory data analysis.

4.1 Elements of exploratory data analysis for insurance data

In order to better understand a relationship between the independent variables and lapse, we did exploratory visualization (mosaic plots, density plots,...) and found out that there is no relationship between gender and lapse in our data. Data also indicates, that there is a difference between lapsed policies of two different types of insurance (a Unit Linked Life insurance and a Traditional Life insurance). This may be caused by the fact, that the unit linked life insurance product is more expensive and less easy understandable than the traditional insurance product. It also seems that policies with quarterly payment have the highest lapse rate. On the other hand, the policies paid with one single payment and policies with monthly payments have the lowest lapse rate. The policies of customers who migrated five times have the largest lapse rate. Generally, lapse contracts have shorter duration, hence they have lower number of migrations.

From our analyses it followed that the policies of customers who are at the age between 20 and 35 at the time of conclusion of the contract are more likely to lapse than the policies of older customers. Moreover, the shorter time elapsed since contract information was updated the lower risk of contract lapse occurs. There is a very similar dependency between lapses and policy duration. The lapse rate is higher for policies with shorter duration.

The region and the number of migrations are relevant behavioural characteristic for lapse prediction only if assumption, that lapse rate is higher in the poorer regions, is cor-

rect. In order to validate this assumption we examine the relationship between lapse rate and selected macro economical aggregates by regions. We have chosen the net disposable income of households per capita, GDP per capita and unemployment rate. Values of income and GDP come from the Czech Statistical Office and the source of values of unemployment rate is the Czech Ministry of Labor and Social Affairs. In Table 3, we summarize selected indicators and add the proportion of people with overdue liabilities to the total population at the age of 18 and over in each region (source: `www.solus.cz`) and proportion of lapse contracts to total contracts per region (source: the insurance company). The highest lapse rate occurs in regions with the highest unemployment rate (Ústecký region 42%) and the lowest disposable income (Liberecký 41%, Karlovarský 40%, Olomoucký 40%). Also the highest rate of people having problem with paying off their debts occurs in the poorer regions (Ústecký 14%, Karlovarský 13%, Liberecký 11%).

Table 3: Selected macro economical aggregates, payment behaviour and lapse rate by region.

Region	Income	GDP	Unempl.	Liab.	Lapse
Capital city Prague	250,121	768,173	0.04	0.06	0.36
Středočeský region	206,669	325,797	0.07	0.08	0.37
Jihomoravský region	184,823	341,024	0.10	0.07	0.37
Královéhradecký region	179,715	315,307	0.08	0.07	0.38
Pardubický region	177,064	297,755	0.08	0.07	0.39
Region Vysočina	180,102	303,263	0.09	0.05	0.39
Zlínský region	178,580	308,642	0.09	0.05	0.39
Moravskoslezský region	176,135	317,835	0.11	0.10	0.39
Jihočeský region	181,215	306,576	0.08	0.07	0.4
Plzeňský region	187,924	326,513	0.07	0.08	0.4
Karlovarský region	171,785	260,083	0.10	0.13	0.4
Olomoucký region	172,415	281,540	0.11	0.07	0.4
Liberecký region	178,750	279,733	0.10	0.11	0.41
Ústecký region	170,925	289,851	0.13	0.14	0.42

Income = Net disposable income of households per capita (in CZK, year 2011),
GDP = GDP per capita (in CZK, year 2011),
Unempl. = Unemployment rate (%, value to date 31.12.2011),
Liab. = Proportion of people with overdue liabilities to the total population (% value to date 31.3.2012),
Lapse = Proportion of lapsed contracts to total contracts per region (%).

Table 4 shows correlation coefficients among indicators. The lapse rate negatively correlates with disposable income (-0.73), i.e. the lower income, the higher lapse rate. We can observe a positive correlation between the lapse rate and two indicators: the unemployment rate (0.67) and payment behaviour (0.58). Consequently, we may assume that the poorer regions of Czech Republic might have a higher risk of lapse.

Our exploratory analysis indicates that gender can be omitted from lapse prediction modelling. This fact can be used also for decision that gender does not need to be collected by the insurance company.

Table 4: Pearson's Correlation Coefficients.

	Income	GDP	Unempl.	Liab.	Lapse
Income	1.00	-	-	-	-
GDP	0.94	1.00	-	-	-
Unempl.	-0.81	-0.69	1.00	-	-
Liab.	-0.34	-0.31	0.58	1.00	-
Lapse	-0.72	-0.63	0.68	0.60	1.00

4.2 Data quality analysis of insurance data

In effort to increase the reliability of our analyses, e.g. lapse prediction modelling, we can choose a suitable set of variables for our basic model not only on the basis of performed exploratory analysis, but also on the basis of data quality. Moreover, if we describe data quality of all available statistical units, we can use this information for data set preparation and for better understanding of a resulting model. If we would like to assess data quality of a contract, it is necessary to start with data quality of individual variables. For simplicity, and in coherence with our statements in the previous sections, in further analysis we restrict ourselves to the three elements of data - currency, timeliness and uncertainty. It is obvious that variables gender, age of a client and type of product are constant values at the time of conclusion of a contract, therefore unimportant for the intendet data quality analysis. We omit them from the rest of our data quality analysis assuming that there are no data issues for these three variables, i.e. the data quality with respect to uncertainty is 1, timeliness and currency are irrelevant for these variables.

Currency and timeliness were already defined in Section 2. But for our purposes it is necessary to transform them to the unit interval in order to get 1 as the best possible option and 0 as the worst one. We use a very simple transformation $\frac{a}{(a+x)}$, where x represents currency or timeliness, respectively, and a represents our sensitivity to changes in data quality with respect to currency and timeliness. For simplicity, presented results were computed for $a = 1$. The extraction date was October 1, 2013 and the delivery date was November 10, 2013. In our case, currency is the same for all variables and also for all statistical units. Timeliness was computed using the same formula for all variables but, in general, it is different for different statistical units.

Uncertainty represents our doubts about data quality of an individual variable. In our opinion, contrary to timeliness and currency, evaluation of uncertainty cannot be entirely based on a particular value of the variable corresponding to the selected contract, but we should primarily evaluate the whole variable. In the case of the presented data set, the variable region is validated by a financial intermediary (an agent, a broker). The insurance company records all characteristics of the contract proposal received from the intermediary into its primary production information system. After that the client receives his or her contract and confirms correctness of information by the act of acceptance, hence all data can be considered reliable at the time of inception. Contract's characteristics are under the insurer's control. The contact address region could be invalid if the client is not motivated to update it after migrating to a different place. Therefore in our example uncertainty is interesting only for the variables region and the number of migrations, data

quality with respect to uncertainty equals to 1 for the rest of variables .

We decided to compute uncertainty of the number of migrations as follows. We suppose that the contract with a positive number of policyholder migrations is correct because the client is rigorous and updates his or her personal data. Similarly, we suppose the correct data for contracts within the three-month period of the confirmation process. For these contracts the uncertainty degree (ud) is 1 $(\mathrm{ud}(\mathrm{contract}_k) = 1, k = 1, 2, \ldots, n_1)$, where n_1 is the number of contracts with a positive number of policyholder migrations or contracts shorter than three months.

The uncertainty degrees for the rest of contracts were determined using the following procedure. The average of the variable number of policyholder migrations was computed for each region. For each remaining contract, the uncertainty degree was computed according to the formula

$$\mathrm{ud}(\mathrm{contract}_k) = \mathrm{P}(0 \text{ migrations in a region}_l), k = 1, 2, \ldots, n_2; l = 1, 2, \ldots s, \quad (4.1)$$

where n_2 is the number of contracts older than three months or with zero policyholder migrations, s is the number of regions and P is the probability mass function of the Poisson probability distribution with the mean λ_l estimated by the mean number of policyholder migrations in the region l. Formula (4.1) is coherent with intuition that migrations are more likely for regions with the higher average number of policyholders migrations. The uncertainty degrees of contracts with respect to the number of policyholder migrations were used also for regions.

Using the above mentioned procedure, we assign to each variable (except variables gender and age) in each contract a triplet (currency, timeliness, uncertainty). Therefore each contract is characterized by a family of triplets, one for each variable, i.e. it is modelled as a hesitant fuzzy set. Then DQI for the contract can be computed as a distance between the contract itself and the ideal contract. In the paper we compute the distance using two basic distances, the normalized Hamming distance (NHD) (3.5) and the normalized Euclidean distance (NED) (3.6). Values of DQI are visualized in Figure 1 in the form of empirical density plots. From Figure 1 it is obvious that, regardless of the metric, the number of contracts with low data quality of selected variables with respect to currency, timeliness and uncertainty is quite high. Using computed DQI we can conclude that due to the low data quality with respect to currency, timeliness and uncertainty, we can expect the low predictive power of resulting models. Moreover, in addition to results of exploratory data analysis on the basis of low data quality, we can exclude variables region and number of migrations from the set of variables assumed as predictors for our lapse models.

As it was already mentioned before, we can use DQI also to compute weights for individual cases. Despite the fact that in our example weighting of cases would not decrease error rates of the resulting lapse prediction models, we prefer to include DQI into the model fitting process because it could boost our confidence in results we got.

5 Discussion and conclusions

The main result of the presented paper is a measure of data quality, so called Data Quality Index (DQI). It allows us to evaluate data quality of individual contracts by a single

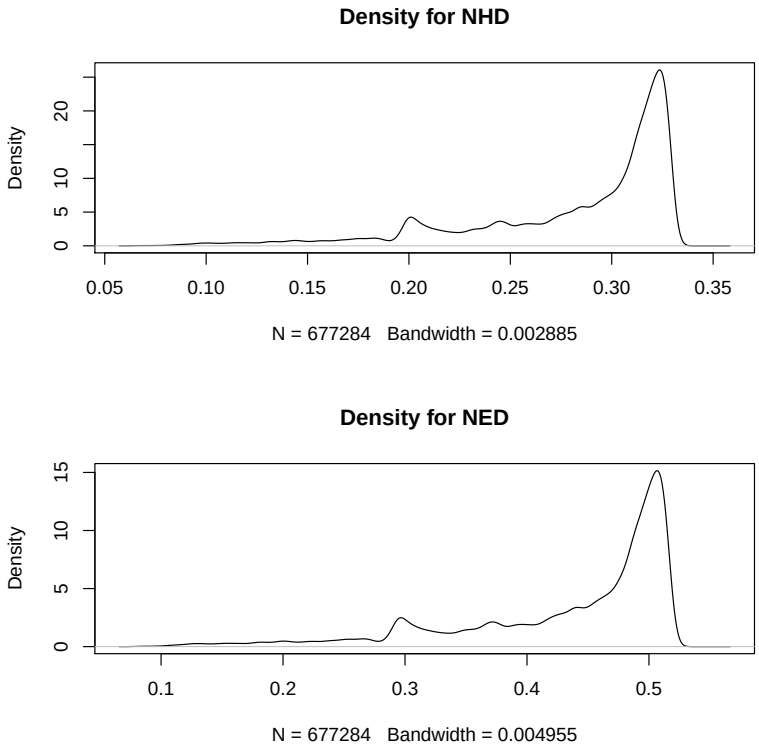

Figure 1: Density plots for DQI computed NHD (upper plot) and NED (lower plot)

number from the unit interval. The starting point of the whole procedure is based on evaluation of data quality of individual variables. The proposed measure is illustrated using the real insurance data set and some possibilities how to incorporate data quality analysis into complex data analysis procedures are pointed out. Although in the paper we restrict ourselves only to currency, timeliness and uncertainty, and we use very simple distances to assess each of them, it is obvious that our analysis can be further extended using more components and more sophisticated mappings for these components. Alternatively, using the same mathematical representation of data quality for individual contracts, we can use an appropriate aggregation function instead of a distance based measure to evaluate data quality. These possible extensions, as well as other behavioural factors (occupation, education) and distances between individual contracts, will be further investigated. Moreover, because all key elements of the presented data quality analysis, such as values of DQI, their interpretation, appropriate data quality components and distances etc., are strictly context dependent, we will focus on DQI in particular contexts in our future research, e.g. on verification of possibility to use DQI as a prior to adjust lapse probability models constructed using some well established classification methods (logistic regression, random forests etc.).

The codes for all the examples given above are written in R (R core team 2013) and are included in supplementary materials of the paper. In order to further simplify possible adoption of the proposed methodology we also included a small artificial data set mimicking some properties of the original one.

Acknowledgment

This work was supported by projects Mobility - enhancing research, science and education at Matej Bel University, ITMS code: 26110230082, under the Operational Program Education co-financed by the European Social Fund, VEGA 1/0647/14 and IGA VSE F4/17/2013.

We would like to thank prof. Hana Řezanková for her valuable comments and suggestions.

References

[1] Berti-Equille, L. (2007): Quality Awareness for Data Management and Mining. *Habilitation a Diriger des Recherches*, Universit'e de Rennes 1, France, [available online]

[2] Bonnissone, P.P. and Tong, M. (1985): Editorial: Reasoning with Uncertainty in Expert Systems, *Int'l J. Marl Machine Studies*, **22**, 241–250.

[3] Bosc, P. and Prade, H. (1993): An Introduction to Fuzzy Set and Possibility Theory Based Approaches to the Treatment of Uncertainty and Imprecision in Database Management Systems,*Proc. Second Workshop Uncertainty Management in Information Systems: From Needs to Solutions*, 44–70, Catalina, Calif.

[4] Cox, D.R. (1972): Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

[5] Grahne, G. (2002): Information Integration and Incomplete Information, *IEEE Data Eng. Bull.*, **25(3)**, 46–52.

[6] Hwang, C. L. and Yoon, K. (1981): *Multiple Attribute Decision Making: Methods and Applications*. New York: Springer-Verlag.

[7] Imielinski, T., Lipski, W.JR. (1984): Incomplete Information in Relational Databases.*J. ACM*, **31(4)**, 76–791.

[8] Motro, A. (1993): Sources of Uncertainty in Information Systems, *Proc. Second Workshop Uncertainty Management and lnformation Systems: From Needs to Solutions*, 9–26, Catalina, Calif.

[9] Naumann, F., Freytag, J.-Ch., Leser, U. (2004): Completeness of Integrated Information Sources. *Inf. Syst.*, **29(7)**, 58–615.

[10] Naumann, F., Leser, U., Freytag, J.-Ch. (1999): Quality Driven Integration of Heterogenous Information Systems, *Proceedings of the 25th International Conference on Very Large Data Bases*, 447–458, Edinburgh, Scotland, UK.

[11] Parsons, S. (1996): Current approaches to handling imperfect information in data and knowledge bases. *IEEE Transactions on Knowledge and Data Engineering*, **8**, 353–372.

[12] Peralta, V. (2006): Data Quality Evaluation in Data Integration Systems. *Ph.D. thesis*, Université de Versailles, France and Universidad de la República, Uruguay.

[13] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[14] Segev, A. and Fang, W. (1990): Currency-Based Updates to Distributed Materialized Views. *Proceedings of the 6th International Conference on Data Engineering*, ICDE 1090, 51–520, Los Angeles, CA, USA.

[15] Theodoratos, D. and Bouzeghoub, M. (1999): Data Currency Quality Factors in Data Warehouse Design, *Proceedings of the International Workshop on Design and Management of Data Warehouses*, DMDW'99, Heidelberg, 15.1–15.16, Germany.

[16] Zeshui Xu and Meimei Xia (2011): Distance and similarity measures for hesitant fuzzy sets, *Information Sciences*, **181**, 2128–2138.