Integrated Computer-Aided Engineering 21 (2014) 355–366
DOI 10.3233/ICA-140473
IOS Press

On the use of fuzzy partitions to protect data

Pelayo Quirós[a], Pedro Alonso[a,*], Irene Díaz[b] and Susana Montes[c]
[a]*Department of Mathematics, University of Oviedo, Oviedo, Spain*
[b]*Department of Computer Science, University of Oviedo, Oviedo, Spain*
[c]*Department of Statistics and O.R., University of Oviedo, Oviedo, Spain*

Abstract. Data protection is one of the most challenging tasks nowadays due to the huge amount of information that can be shared and crossed from different sources. Releasing microdata is a way to protect data, mainly in the economic and medical field. However, this kind of data can experience privacy attacks. This paper proposes the use of fuzzy sets as a way to improve the protection of privacy in microdata. Then, traditional definitions of k-anonymity, l-diversity and t-closeness are extended. The performance of these new approaches is checked in terms of the risk index.

Keywords: k-anonymity, l-diversity, t-closeness, fuzzy partition, data privacy

1. Introduction

Nowadays, with the communications development, a big amount of information is shared all around the world, and some of this information is released as microdata (i.e., data not summed up in statistics, but directly related to individuals), mainly in the economic and medical field. Microdata are often presented by tables containing information about individuals. With the purpose of avoiding individuals to be uniquely identified, a common practice for organizations is to remove explicit identifiers such as name, telephone or social security number. However, although sometimes the published table looks anonymous, the privacy of the releasing data is unintentionally compromised. For example, joining the data available in a released table with some publicly accessible database (like `Census` database) or other attributes (for example race or ZipCode) can be used to identify individuals. The problem of protecting private information is actually legislated in several countries. The most representative laws regulating this task are the United States Healthcare Information Portability and Accountability Act and European Union directive 95/46/EC.

Therefore, as some public administrations are required to make public certain information, there is a need to find the balance between the right to privacy and the data dissemination. In order to avoid the identification of individuals, some techniques have been developed to avoid this problem. Most of them are based on grouping individuals into equivalence classes, as the well known k-anonymity. Samarati [35] and Sweeney et al. [40] define this technique as the property that makes every individual in the data indistinguishable from al least other $k-1$ individuals. There are many techniques to efficiently achieve k-anonymous tables efficiently, as the ones proposed by [35,40], or some more elaborated techniques, as Matatov's et al. one (see [28]).

However, k-anonymity does not represent a good protection if sensitive values in an equivalence class lack diversity (homogeneity attack). A new technique was developed by Machanavajjhala et al. (see [27]) called l-diversity, that requires the sensitive attributes in each group of k indistinguishable individuals to have at least l well represented different values.

The l-diversity presents also some drawbacks, mainly based on bias and similarity attacks. For instance, if the sensitive attribute is numeric, the l-diversity does not take into account that some values can be very

*Corresponding author: Pedro Alonso, Department of Mathematics, University of Oviedo, Avda. Calvo Sotelo s/n, Oviedo 33007, Spain. Tel.: +34 985 182 128; Fax: +34 985 103 354; E-mail: palonso@uniovi.es.

ISSN 1069-2509/14/$27.50

similar. To solve these similarity attacks, a new technique was recently developed by Li et al. [24], known as t-closeness, which establishes that the distribution of the sensitive attribute in each equivalence class has to be similar to the one in the whole table. In this approach, the similarity is measured by means of the Earth Mover's Distance (see [24]).

There are other approaches to preserve privacy in data contexts. For example, [6] it is introduced local suppression to achieve a tailored privacy model for trajectory data anonymization. Zhong [48] studies how to maintain privacy in distributed mining of frequent itemsets without revealing each party's portion of the data to the other. Other highlighted works about this issue can be seen [12,13,36,37].

Each technique has a weak point. The proposal made in this paper is based on the use of the properties of fuzzy sets to improve these techniques and to get a better protection. Therefore, each attribute will be masked by fuzzy partitions instead of the crisp ones. This is not a simple generalization, as it requires a different way to count elements (cardinality) (see [11]). Therefore, it becomes necessary to analyze the different definitions of cardinality for a fuzzy set, that can produce an scalar or a fuzzy number. The first ones associate to each fuzzy set a quantity, while the fuzzy definitions associate to each fuzzy set a function on the unit interval. As it will be justified later, the approximation used in this work will be the second one.

In addition, fuzzy sets provide a natural way of grouping data by a different level of membership. In this way, the similarity between data can be expressed by the difference between the membership degree of data to a given fuzzy set. This can be assumed for both mono- and multi-dimensional data. There are many works related to fuzzy theory as a tool for solving engineering problems (see, for example, [2,33]). The goal of this work is to protect released data using fuzzy set theory and to check the performance of this approach. There are some works related to the study of privacy using fuzzy techniques (see [10,43]).

The goodness of using fuzzy sets has already been demonstrated in a large number of papers, where they are related to fields like engineering (see [1,14,20,32, 45,46]). In particular, in relation to information processing in natural and artificial neural systems, some works are specially interesting, like [5,22,26,38,39, 42].

Other situations where these techniques are employed to act as a bridge between advances being made in computer technology and civil and infrastructure engineering can be seen [9,16,18,23,44,47].

The main goal of this paper is to extend the three aforementioned privacy techniques to a fuzzy environment, where the improvements that this generalization provide are shown in a experimentation carried out with two different real data sets.

The paper is organized as follows. Section 2 introduces the definition of k-anonymity, l-diversity and t-closeness as well as it describes some concepts about fuzzy sets related to this work. In Section 3 the proposal of this work is presented. Section 4 is devoted to check the performance of the approach with real microdata. `Census` and `Tarragona` datasets will be used for that. The conclusions are presented in last section.

2. Basic concepts

Basic notions and notations concerning privacy metrics and fuzzy sets used throughout the paper are introduced in this section.

2.1. Privacy metrics

This subsection briefly describes the methods and tools related to privacy upon which this work is based. For a more detailed description see [7,25,27].

Microdata are represented by tables, where the rows represent the individuals, and the columns the attributes defining the individuals. In a privacy context, two types of attributes are defined, the sensitive ones (the ones to be protected, denoted by S), and the non sensitive ones (the others, denoted by Q). A quasi-identifier is a subset of the non sensitive attributes.

Given a quasi-identifier $QI = \{A_1, \ldots, A_n\}$, a partition of *QI* is given by the combination of the partitions of each attribute A_i. Partitions are used to mask microdata in order to minimize the risk of revealing private information. This new table, the protected one, groups individuals with the same values with regard to the quasi-identifier. These blocks will be called hereinafter q^*-blocks.

In the remaining part of this subsection, three well known techniques to protect private information in microdata (k-anonymity, l-diversity and t-closeness) are explained.

Samarati and Sweeney (see [35,40]) define a k-anonymous table as the one that makes every individual in the data indistinguishable from al least other $k-1$ individuals. The formal definition is given below.

Definition 1. A table T satisfies k-anonymity if for all tuple $t \in T$, it exists other $k-1$ tuples indistinguishable

respect to the quasi-identifier, i.e., it exists $t_{i_1}, \ldots, t_{i_{k-1}} \in T$ such that $t[QI] = t_{i_1}[QI] = \ldots = t_{i_{k-1}}[QI]$, where $t[QI]$ denotes the values taken by the tuple t for the quasi-identifier *QI*.

If the values of the sensitive attribute in one block are the same, the data are unprotected against a homogeneity attack. To prevent this attack a new metric is developed. According to [27], l-diversity is based on the following concepts.

Definition 2. A q^*-block is l-diverse if it contains at least l well-represented values for the sensitive attribute S. A table is l-diverse if every q^*-block is l-diverse.

To measure whether a q^*-block is l-diverse, it is necessary to obtain the posterior belief through a generalization T^*. Definition 3 states how it is computed.

Definition 3. Let q be a value of the non sensitive attribute Q in the table T; let q^* be the generalized value of q in the private table T^*; let s be a possible value of the sensitive attribute; let n(q*,s') be the number of tuples $t^* \in T^*$ where $t^*[Q] = q^*$ and $t^*[S] = s'$; and let $f(s'|q^*)$ be the conditional probability of the sensitive attribute conditioned on the fact that the non sensitive attribute Q can be generalized to q^*. Then the posterior belief is defined by:

$$\beta_{(q,s,T^*)} = \frac{n_{(q^*,s)} \dfrac{f(s|q)}{f(s|q^*)}}{\displaystyle\sum_{s' \in S} n_{(q^*,s')} \frac{f(s'|q)}{f(s'|q^*)}}.$$

Even if a table satisfies l-diversity, it could be attacked using similarity of values. If all the possible sensitive values in a block are similar or closely related, the attacker can get some undesirable extra information. The t-closeness (see [24]) tries to prevent this attack.

Suppose an attacker has a prior belief about the sensitive attribute of an individual (denoted by B_0). He also gets the information about the whole published table (denoted by W). Then his belief changes to B_1. After identifying the values of the quasi-identifier, the attacker identifies the block the individual belongs to, and it will get the distribution of the sensitive attribute in that block (denoted by P). Then, the attacker's belief changes to B_2.

$$B_0 \xrightarrow{W} B_1 \xrightarrow{P} B_2$$

Assuming that the information given by W is public, if the goal is to minimize $B_2 - B_0$, $B_2 - B_1$ should be minimized. To measure that distance the Earth Mover's Distance (see [24]) was used. This distance only takes into account the order of the values and it does not consider the similarity between them. The following two definitions formalize these concepts

Table 1
Original data table

Individual	Age	Fnlwgt	Hours per week
1	39	77516	40
2	39	83311	13
3	38	215646	40
4	53	234721	40
5	28	338409	40
6	37	284582	40
7	49	160187	16
8	52	209642	45
9	31	45781	50
10	42	159669	40
11	37	280464	80
12	30	141297	40
13	30	141297	60
14	30	141297	80
15	34	245487	45

Definition 4. Let $\{v_1, \ldots, v_m\}$ be the numeric values assumed by the sensitive attribute, which are ordered increasingly. Let $P = (p_1, \ldots, p_m)$ be the distribution of the sensitive attribute in the block and $W = (w_1, \ldots, w_m)$ be the distribution of the sensitive attribute in the whole table. The Earth Mover's Distance is given by

$$D[P, W] = \frac{1}{m-1} \sum_{i=1}^{m} \left| \sum_{j=1}^{i} (p_j - w_j) \right|,$$

where $v_i \leqslant v_j$ if $i \leqslant j$ for quantitative attributes, and

$$D[P, W] = \frac{1}{2} \sum_{i=1}^{m} |p_i - w_i|,$$

for qualitative attributes.

Based on this distance, the definition of t-closeness is given as follows.

Definition 5. A block is said to satisfy t-closeness if the distance between the distribution of a sensitive attribute in this class P and the distribution of the attribute in the whole table W is no more than a threshold t. A table is said to have t-closeness if all blocks have t-closeness.

Example In Table 1, 15 individuals from the *Adult data set from the UC Irvine Machine Learning Repository* [3] are shown, with the values associated to the

Table 2
Table satisfying 2-anonymity, 2-diversity and 0.1111-closeness

Age	Fnlwgt	Hours per week
$(0, 35]$	I_A	40
$(0, 35]$	I_A	50
$(0, 35]$	I_A	60
$(0, 35]$	I_A	80
$(0, 35]$	I_B	40
$(0, 35]$	I_B	45
$(35, 40]$	I_A	13
$(35, 40]$	I_A	40
$(35, 40]$	I_B	40
$(35, 40]$	I_B	40
$(35, 40]$	I_B	80
$(40, \infty)$	I_A	16
$(40, \infty)$	I_A	40
$(40, \infty)$	I_B	40
$(40, \infty)$	I_B	45

attributes *Age, fnlwgt* and *hours per week*, where the latter is the one treated as sensitive.

Let us apply a partition for each attribute of the non sensitive ones, *Age* and *fnlwgt*. The one applied for *Age* is (0,35], (35,40] and (40,∞). The second attribute is partitioned in two sets, I_A = [0,200000) and I_B = [200000,∞). Table 2 is the obtained one.

There are 6 q^*-blocks in the table, where the smallest ones has 2 individuals, hence, it is a 2-anonymous table. In each q^*-block, there are at least two different values of the sensitive attribute, so the table is 2-diverse.

Regarding the study of the t-closeness, given Q = {13,16,40,45,50,60,80} the set of values that the sensitive attribute take, the distributions in the whole table (W) and in each q^*-block ($P_1, \ldots, P_6$) are given by

$$W = \left(\frac{1}{15}, \frac{1}{15}, \frac{7}{15}, \frac{2}{15}, \frac{1}{15}, \frac{1}{15}, \frac{2}{15}\right),$$

$$P_1 = \left(0, 0, \frac{1}{4}, 0, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right),$$

$$P_2 = \left(0, 0, \frac{1}{2}, \frac{1}{2}, 0, 0, 0\right),$$

$$P_3 = \left(\frac{1}{2}, 0, \frac{1}{2}, 0, 0, 0, 0\right),$$

$$P_4 = \left(0, 0, \frac{2}{3}, 0, 0, 0, \frac{1}{3}\right),$$

$$P_5 = \left(0, \frac{1}{2}, \frac{1}{2}, 0, 0, 0, 0\right),$$

$$P_6 = \left(0, 0, \frac{1}{2}, \frac{1}{2}, 0, 0, 0\right).$$

Applying the Earth Mover's Distance for each q^*-block, the obtained values are $D[P_1, W] = 0.2417$, $D[P_2, W] = 0.15$, $D[P_3, W] = 0.3$, $D[P_4, W] = 0.1111$, $D[P_5, W] = 0.2389$ and $D[P_6, W] = 0.15$. Hence, $t = \min(D[P_i, W] | i \in \{1, \ldots, 6\}) = 0.1111$.

In summary, each technique has its own drawbacks: k-anonymity suffers from homogeneity attacks, l-diversity from the similarity of values in the sensitive attribute and t-closeness is unintuitive, as well as the additional information which is also troublesome. The new techniques that are developed in the next section using fuzzy partitions, provide a good way to palliate these problems, thanks to the good properties of the fuzzy sets.

2.2. Fuzzy cardinality

In a k-anonymous table the equivalence classes are constructed so that each individual record is indistinguishable from at least $k - 1$ records within the same class (with regard to the quasi-identifier). This condition might not be enough, as sensitive information could be homogeneously associated to individuals within the same class. This drawback could be overcomed by introducing a fuzzy model. A priori, the fuzzyfied released table provides a first level of privacy because some uncertainty is introduced to protect the data against an attacker and at the same time it is obtained a more informative released table.

However, the introduction of fuzziness requires a redefinition in terms of fuzzy sets of the aforedefined metrics.

Therefore, if $X = \{x_1, x_2, \ldots, x_n\}$ is the universe of discourse, a fuzzy set $A \subset X$ is characterized by its membership function $\mu_A(x) : X \to [0, 1]$. If the membership of the corresponding fuzzy set represents a degree of uncertainty, the membership of an element of the referential, $x_i \in X$, to a fuzzy set A means the degree of possibility that an imprecisely known parameter x_0 has value x_i, when the available information about x_0 is that "x_0 is A" (see [8]). Note that $\Pi : \mathcal{P}(X) \to [0, 1]$ is a possibility measure if there exists a map $\pi : X \to [0, 1]$ such that

$$\Pi(A) = \sup_{x \in A} \pi(x), \forall A \subseteq X,$$

where π is called possibility distribution.

Returning to the initial definition of fuzzy sets, the simplest way to count its elements is the σ-count which is defined as follows:

$$|A|_\sigma = \sum_{x \in X} \mu_A(x).$$

However, there are many ways of counting fuzzy sets. In this work it is used the concept of fuzzy cardinality developed in [31] where the cardinality of a fuzzy set is defined to be k to a certain degree. According to this approach the fuzzy cardinality of a fuzzy set A is defined by

$$|A|_f(k) = \mu_{(k)} \wedge (1 - \mu_{(k+1)}), k = 0, \ldots, n, \quad (1)$$

where $\mu_{(1)}, \mu_{(2)}, \ldots, \mu_{(n)}$ represent the values of $\mu_A(x_1), \mu_A(x_2), \ldots, \mu_A(x_n)$ arranged in a decreasing order of magnitude and $\mu_{(0)} = 1, \mu_{(n+1)} = 0$. Note that $|A|_f(k) = \mathit{Poss}(|A| = k)$ (see [31]), where *Poss* denotes a possibility measure.

This cardinality is chosen among others as it is the one which fits the best with the aims of our proposal, due to the fact of being defined to take a value to a certain degree.

From Eq. (1) it is concluded (see [31]) that the fuzzy cardinality $|A|_f$ is a fuzzy convex set and the possibility that A has at least k elements is:

$$\mathit{Poss}(|A|_f \geqslant k) = \begin{cases} \mu_{(k)}, & \text{if } k \geqslant j, \\ (1 - \mu_{(j)}) \vee \mu_{(j)}, & \text{if } k < j, \end{cases}$$

where

$$j = \begin{cases} \max\{1 \leqslant s \leqslant n \mid \mu_{(s-1)} + \mu_{(s)} > 1\}, & \text{if } A \neq \emptyset, \\ 0, & \text{if } A = \emptyset. \end{cases}$$

Thus, the non fuzzy cardinality for a fuzzy set A is defined as follows:

$$|A|_{nf} = \begin{cases} 0, & \text{if } A = \emptyset, \\ j, & \text{if } A \neq \emptyset \text{ and } \mu_{(j)} \geqslant 0.5, \\ j-1, & \text{if } A \neq \emptyset \text{ and } \mu_{(j)} < 0.5. \end{cases} \quad (2)$$

This definition of cardinality is used to define a new concept of k-anonymity based in a different way of counting, assuming the nature of the information to be fuzzy. In addition to cardinality, a very important concept in this context is the concept of partition. In this case, partitions formed by fuzzy sets are used. There are several different approaches in the literature (see, for instance, [30]), but the classical definition given by Ruspini [34] is considered in this work, since it is the most used in applied fields by its simplicity for computation.

Definition 6. Let A be an attribute of the table T and $D(A)$ the different values that the attribute can take. A fuzzy partition of this attribute is given by a family of fuzzy sets $\{A_1, \ldots, A_n\}$ such that:

$$\forall x \in D(A), \sum_{i \in \{1,\ldots,n\}} \mu_{A_i}(x) = 1,$$

where μ_{A_i} denotes the membership function of A_i for any $i \in \{1, \ldots, n\}$.

As each variable in the quasi-identifier (*QI*) is masked according to a fuzzy partition, it will be called fuzzy quasi-identifier (*FQI*).

3. Privacy techniques based on fuzzy partitions

In this section the effect of introducing fuzzy partitions to protect a released table is studied. The three techniques which were given in the previous section (k-anonymity, l-diversity and t-closeness), are defined when fuzzy partitions are used instead of the crisp ones, providing a better way to deal with the attacks previously explained.

First, it is introduced a new privacy metric based on assigning a fuzzy number to anonymity (for more information authors refer to [11]). Secondly, it is checked that the traditional privacy metrics l-diversity and t-closeness can be extended using fuzzy partitions.

3.1. Q-anonymity

A method whose properties are pretty similar to k-anonymity, called Q-anonymity is defined in this subsection.

Definition 7. Let T be a table with attributes $\{A_1, \ldots, A_n\}$, and let *FQI* be a fuzzy quasi-identifier associated with it. The Q-anonymity of T with respect to *FQI* is given by the possibility distribution:

$$\mathit{Poss}(|T|_f \geqslant Q) = \mathcal{T}(\beta_{f_1}, \ldots, \beta_{f_s}),$$

being β_{f_i} the possibility that $T[FQ]_i$ has at least Q elements and $T[FQI]_1, \ldots, T[FQI]_s$ the different fuzzy classes. *Poss* is a possibility measure and $\mathcal{T}$ is an aggregation operator.

Note that an aggregation operator (see, for instance, [15]) is a map $\mathcal{T}$: $\cup_{n \in \mathbb{N}}[0,1]^n \to [0,1]$ fulfilling the boundary conditions ($\mathcal{T}(0, \ldots, 0) = 0$ and $\mathcal{T}(1, \ldots, 1) = 1$), being the identity when unary ($\mathcal{T}(x) = x, \forall x \in [0,1]$) and being increasing ($\forall n \in \mathbb{N}$: $x_1 \leqslant y_1, \ldots, x_n \leqslant y_n \Rightarrow \mathcal{T}(x_1, \ldots, x_n) \leqslant \mathcal{T}(y_1, \ldots, y_n)$).

3.2. l-diversity

Definition 3 about the posterior belief is adapted to introduce fuzzy partitions.

First, the term $n(s, q, T^*)$ is redefined according to the next definition. Note that given a fuzzy set A, $\#A$ denotes the cardinality given by Eq. (2).

Definition 8. Let $s \in S$ and $q \in Q$, where S and Q denote the sensitive and non sensitive attributes respectively. Let us suppose that the fuzzy partition of Q is given by the sets $\{Q_1, \ldots, Q_n\}$, and let us denote by $n(s, q, T^*)$ the number of elements in the published table T^* with values s and q for the sensitive and non sensitive attributes respectively. Then

$$n(s, q, T^*) = \sum_{i=1}^{n} \mu_{Q_i}(q) \cdot \#(Q_i \cap s).$$

Secondly, the conditional probability of the sensitive attribute given the generalization of the non sensitive to q^* ($f(s|q^*)$) is redefined according to the following definition.

Definition 9. Let $s \in S$ and $q \in Q$, where S and Q denote the sensitive and non sensitive attributes respectively. Let us suppose that the fuzzy partition of Q is given by the sets $\{Q_1, \ldots, Q_n\}$, then

$$f(s|q^*) = \sum_{i=1}^{r} \mu_{Q_i}(q) \cdot f(s|Q_i),$$

where, $\forall i \in \{1, \ldots, n\}$,

$$f(s|Q_i) = \frac{f(s \cap Q_i)}{f(Q_i)} = \frac{\#(s \cap Q_i)}{\#Q_i}. \tag{3}$$

As it happened in the case of crisp partitions, $f(s|q^*)$ is a probability as the following result states.

Theorem 10. Let $s \in S$ and $q \in Q$, where S and Q denote the sensitive and non sensitive attributes respectively and let $\{Q_1, \ldots, Q_n\}$ be a fuzzy partition of Q. For any $i \in \{1, 2, \ldots, n\}$, the function $f(\cdot|Q_i)$ defined in Eq. (3) is a probability.

Proof It is immediate, by the definition of cardinality, that it is positive and assumes the value 1 in S. If $(A_n)_n \subset S$, such that $A_n \cap A_m = \emptyset$, $\forall n, m$, then:

$$\begin{aligned} f(\cup_n A_n|Q_i) &= \frac{\#\left(\left(\bigcup_n A_n\right) \cap Q_i\right)}{\#Q_i} \\ &= \frac{\#\left(\bigcup_n (A_n \cap Q_i)\right)}{\#Q_i} \\ &= \frac{\sum_n \#(A_n \cap Q_i)}{\#Q_i} \\ &= \sum_n \frac{\#(A_n \cap Q_i)}{\#Q_i} \\ &= \sum_n f(A_n|Q_i). \end{aligned}$$

where the intersection and union of fuzzy sets is given by any t-norm and t-conorm (see [21]), respectively. □

Based on the previous results, the posterior belief is defined as follows.

Definition 11. Let $s \in S$ and $q \in Q$, where S and Q denote the sensitive and non sensitive attributes respectively. Let T^* the published table for the fuzzy partition of Q given by the sets $\{Q_1, \ldots, Q_n\}$. Then

$$\beta_{(q,s,T^*)} = \frac{\left(\sum_{i=1}^{n} \mu_{Q_i}(q) \cdot \#(Q_i \cap s)\right) \gamma_s}{\sum_{s' \in S} \left(\sum_{i=1}^{n} \mu_{Q_i}(q) \cdot \#(Q_i \cap s')\right) \gamma_{s'}},$$

with

$$\gamma_s = \frac{f(s|q)}{\sum_{i=1}^{n} \mu_{Q_i}(q) \cdot \frac{\#(Q_i \cap s)}{\#Q_i}}$$

and

$$\gamma_{s'} = \frac{f(s'|q)}{\sum_{i=1}^{n} \mu_{Q_i}(q) \cdot \frac{\#(Q_i \cap s')}{\#Q_i}}.$$

Therefore, the definition of l-diversity is consistent when using fuzzy partitions, only adapting the corresponding posterior believes to the case of using fuzzy sets.

3.3. t-closeness

Finally, the consistence of t-closeness is checked when fuzzy partitions are used. Remember that t-closeness minimizes the distance between the distribution of the sensitive attribute in the whole table (W) and the distribution of the sensitive attribute associated to each individual in the table.

The distribution associated to each individual when data are fuzzy partitioned is a probability distribution according to the next result.

Theorem 12. Let $A_1, \ldots, A_n$ be the non sensitive attributes and S be the sensitive one, where $s_1, \ldots, s_m$ are the values that it can assume. Given $\{Q_1, \ldots, Q_r\}$ a fuzzy partition, $Q = (q_1, \ldots, q_m)$ the distribution of the sensitive attribute in the whole table, $P_1, \ldots, P_r$ the distribution in $Q_1, \ldots, Q_r$ respectively, with $P_i = (p_1^i, \ldots, p_m^i), \forall i = 1, \ldots, r$, and μ_{Q_i} the membership function of Q_i for each i. Then, for all x individual, $P = \sum_{i=1}^{r} \mu_{Q_i}(x) \cdot P_i$ is a probability distribution.

Proof Let $P = (p_1, \ldots, p_m)$ be. To check that P is a probability distribution, it is proved that P is non negative ($p_i \geqslant 0$) and $\sum_{i=1}^{m} p_i = 1$.

- $p_i \geqslant 0, \quad \forall i = 1, \ldots, m$:

  As $\mu_{Q_i}(x) \geqslant 0$ and $p_j^i \geqslant 0, \forall i, j$. Then:

$$p_j = \sum_{i=1}^{r} \mu_{Q_i}(x) \cdot p_j^i \geqslant 0, \quad \forall j = 1, \ldots, m.$$

- $\sum_{i=1}^{m} p_i = 1$:

  As p_j are probability distributions, $\sum_{i=1}^{m} p_i^j = 1$ and by the definition of fuzzy partition, $\sum_{i=1}^{r} \mu_{Q_i}(x) = 1$. Then:

$$\begin{aligned}\sum_{j=1}^{m} p_j &= \sum_{j=1}^{m} \sum_{i=1}^{r} \mu_{Q_i}(x) \cdot p_j^i \\ &= \sum_{i=1}^{r} \sum_{j=1}^{m} \mu_{Q_i}(x) \cdot p_j^i \\ &= \sum_{i=1}^{r} \mu_{Q_i}(x) \cdot \left(\sum_{j=1}^{m} p_j^i \right) \\ &= \sum_{i=1}^{r} \mu_{Q_i}(x) = 1.\end{aligned}$$

Therefore P is a probability distribution. □

Finally, t-closeness is defined as follows:

Definition 13. An individual satisfies t-closeness if the distance between the distribution of the sensitive attribute associated to the individual P (see Theorem 12) and the distribution of the sensitive attribute in the whole table W is no more than a threshold t. A table is said to have t-closeness if every individual satisfies that property.

Example Let us take into account the original data from Table 1. A partition for each non sensitive attribute (*Age* and *fnlwgt*) is applied. For attribute *fnlwgt*, the same crisp partition as in the other example ($I_A = [0, 200000)$ and $I_B = [200000, \infty)$) is used. For the attribute *Age* a fuzzy partition is selected, which is defined by the sets $A_A = (-\infty, 33, 37)$, $A_B = (33, 37, 38, 42)$, $A_C = (38, 42, \infty)$, where A_A and A_C are triangular fuzzy sets and A_B is a trapezoidal fuzzy set. Table 3 is the resulting one.

Table 3
Protected table by a fuzzy partition

Age	Fnlwgt	Hours per week
A_A	I_A	40
		50
		60
		80
A_A	I_B	40
		45
A_B	I_A	13
		40
A_B	I_B	40
		45
		80
A_C	I_A	13
		16
		40
A_C	I_B	40
		45

Table 4
Possibility measures to get the Q-anonymity

	$k=1$	$k=2$	$k=3$
$Poss(A_A \times I_A)$	1	1	1
$Poss(A_A \times I_B)$	0.75	0.75	0
$Poss(A_B \times I_A)$	0.75	0.75	0
$Poss(A_B \times I_B)$	0.75	0.75	0.75
$Poss(A_C \times I_A)$	0.75	0.75	0.25
$Poss(A_C \times I_B)$	1	1	0
$Poss(\lvert T\rvert_f \geqslant k)$	0.833	0.833	0.333

Firstly, let us analyze the Q-anonymity, which is given by the results shown in Table 4. These results lead us to know that the studied table satisfy Q-anonymity for $Q = 2$. It is straightforward from Table 3 that it also satisfies the fuzzy l-diversity with $l = 2$.

In order to get the parameter for fuzzy t-closeness, the distributions of the sensitive attribute associated to each individual must be obtained, given by:

$$P_1 = P_2 = 0.75P_{BA} + 0.25P_{CA},$$

$$P_3 = P_6 = P_{12} = P_{BB}, \quad P_4 = P_8 = P_{CB},$$

$$P_5 = P_{AB}, \quad P_7 = P_{10} = P_{CA},$$

$$P_9 = P_{12} = P_{13} = P_{14} = P_{AA},$$

$$P_{15} = 0.75P_{AB} + 0.25P_{BB}.$$

where $P_{AA}, P_{AB}, P_{BA}, P_{BB}, P_{CA}, P_{CB}$ are the distributions associated to each set of the fuzzy partition. Finally, it must be calculated the Earth Mover's Distance of each individual to the distribution in the whole table (W), and select the minimum one as the parameter, i.e., $t = \min(D[P_i, W] | i \in \{1, \ldots, 15\}) = 0.1014$.

These three generalizated techniques provide an improved protection to the private information of the individuals. This is mainly caused by the uncertainty that the fuzzy partitions provide, as they do not give such an explicit information of each feature as the crisp partitions, which makes some types of attack much harder to be carried out.

4. Experiments

In this section the performance of the proposed approach in terms of privacy preservation is checked. The level of protection obtained when a database is coded using either a fuzzy partition or a crisp one is studied.

4.1. Analyzed methods

In order to get a crisp partition, the k-means (see [29]) method is used. The algorithm describing the method is summarized as:

Step 1. Randomly place k points representing initial group centroids.
Step 2. Assign each object to the group that has the closest centroid.
Step 3. Recalculate the positions of the k centroids after assigning all objects.
Step 4. Repeat Steps 2 and 3 until the centroids no longer move.

The *Matlab* implementation of this method was used in this experiment (see [29]).

On the other hand, the methods used to obtain fuzzy partitions have been fuzzy c-means and Gustafson-Kessel method. Fuzzy c-means (see [17]) minimize the functional:

$$J(X; U, V) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m ||x_k - v_i||_A^2,$$

where X is the data set, $U = [\mu_{ik}]$ the membership matrix of each individual to each set, $V = [v_1, \ldots, v_c]$ the set of centroids and m a parameter controlling the fuzziness.

Table 5
Summary of results for `Census` database. Each value represents the percentage of best results obtained by each method with regard to each metric

		k-means	Gustafson-kessel	Fuzzy c-means
k-anonymity		25%	18%	*57%*
Q-anonymity		–	14%	*86%*
PTOTVAL	l-diversity	*53%*	16%	31%
	t-closeness	0%	*78%*	22%
TAXINC	l-diversity	*51%*	30%	48%
	t-closeness	0	*78%*	22%

The norm used was:

$$D_{ikA}^2 = ||x_k - v_i||_A^2 = (x_k - v_i)^T A (x_k - v_i),$$

the one induced by $A = I$. The implementation of the used algorithm was $FCMclust$ of the package *Fuzzy Clustering and Data Analysis Toolbox* of *Matlab* (see [4]).

Gustafson-Kessel (see [17]) method extends the standard fuzzy c-means algorithm by employing an adaptive distance norm, in order to detect clusters of different geometrical shapes in one data set. Each cluster has its own norm-inducing matrix A_i, which yields the following inner-product norm:

$$D_{ikA_i}^2 = ||x_k - v_i||_{A_i}^2 = (x_k - v_i)^T A_i (x_k - v_i).$$

The matrices A_i represent optimization variables in the c-means functional, thus allowing each cluster to adapt the distance norm to the local topological structure of the data. The objective functional of the Gustafson-Kessel algorithm is defined by

$$J(X; U, V, A) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m D_{ikA_i}^2.$$

This algorithm is known as $GKclust$ in the aforementioned *Matlab* package (see [4]).

The main difference between the sets generated by both fuzzy methods is that the second one, Gustafson-Kessel, provides sets with more irregular shapes than the fuzzy c-means method, due to the use of various matrices in the definition of this method instead of one.

4.2. Results

Results have been obtained for two different databases, `Census` and `Tarragona`, both available at *sdcMicro* R-package (see [41]), in order to test the performance of the proposed approach.

Table 6
Averages and distances to the optimal values for `Census` database

Criteria		*k-means*		*Gustafson-kessel*		*Fuzzy c-means*	
		$\overline{X}$	D	$\overline{X}$	D	$\overline{X}$	D
k-anonymity		213.62	100.27	200.78	107.05	*278.42*	*23.78*
Measure_risk		164.42	189.36	236.78	257.83	*109.22*	*107.9*
Q-anonymity		–	–	184.36	99.3	*262.29*	*47.63*
PTOTVAL	l-diversity	97.73	41.88	89.55	32.54	*109.56*	*10.7*
	t-closeness	0.0867	0.0827	*0.0049*	*0.0042*	0.0129	0.0115
TAXINC	l-diversity	93.76	37.42	86.13	30.96	*104.6*	*10.19*
	t-closeness	0.0879	0.0841	*0.0047*	*0.004*	0.0122	0.0107

Table 7
Summary of results for `Tarragona` database. Each value represents the percentage of best results obtained by each method with regard to each metric

		k-means	*Gustafson-kessel*	*Fuzzy c-means*
k-anonymity		21%	*55%*	24%
Q-anonymity		–	*72%*	28%
GROSS.PROFIT	l-diversity	22%	*64%*	14%
	t-closeness	0%	*51%*	49%
NET.PROFIT	l-diversity	23%	*65%*	12%
	t-closeness	0%	38%	*62%*

The used quasi-identifiers in both databases are every combination of two elements of the attributes, which are clustered according to the previously defined methods. According to the size of the dataset and in order to better compare the results, the quasi-identifier is coded using three sets (fuzzy or crisp).

After coding the quasi-identifier according to the three algorithms, k-anonymity, l-diversity, t-closeness and Q-anonymity are computed. Note that when fuzzy partitions are considered, the k-anonymity is computed taking as k the estimation got in Eq. (2).

As a complementary check, once the data are protected, the number of individuals with risk of re-identification higher than the rest is computed. This number is computed using the function *measure_risk* of package *sdcMicro* in R (see [41]). This measure of individuals in risk is computed as follows (see [19]):

- For each individual in the released table $i^* \in T^*$, it is computed the probability of this individual to be related to an individual in the original table (ρ_i).
- The individual risk of re-identification, r_i, that represents the same probability of ρ_i, but with the condition that the attacker tries to obtain the values of all the individuals of the released table.
- The output argument is the number of individuals of the table whose r_i is much bigger than the rest.

In the next two subsections, description and obtained results for each database are given and analyzed.

4.2.1. `Census` database

`Census` dataset was obtained on July 27, 2000 using the public Data Extraction System of the U.S. Bureau of the `Census`. It consists of 1080 examples characterized by 13 attributes (*afnlwgt, agi, emcontrb, ernval, fedtax, fica, intval, pearnval, pothval, ptotval, statetax, taxinc* and *walval*).

To test the performance of the approach the attributes *ptotval* (total person income) and *taxinc* (taxable income amount) have been selected as sensitive variable. Therefore two different experiments are considered, one with *ptotval* as sensitive value and other with *taxinc*.

Table 5 shows the number of experiments with the highest level of protection according to each metric and clustering method. As it can be seen in this table, fuzzy c-means performs the best with regard to both k-anonymity and Q-anonymity because it is the method that reaches the highest values for k and Q. Anyhow, this behaviour does not remains when both l-diversity and t-closeness are studied. Focusing on l-diversity, crisp methods seem to perform better. On the other side, Gustafson-Kessel algorithm performs the best with regard to t-closeness. The statistical signification of these comments can be proven by means of the proportions test. Thus, the proportion of times that Q-anonymity combined with fuzzy c-means was the best method is significantly greater than the same proportion for other methods (p-value of the proportion test $4.66 \cdot 10^{-8}$) and the same happens for t-

Table 8
Averages and distances to the optimal values for Tarragona database

Criteria		*k-means*		*Gustafson-kessel*		*Fuzzy c-means*	
		$\overline{X}$	D	$\overline{X}$	D	$\overline{X}$	D
k-anonymity		9.79	13.76	*18.82*	*3.76*	12.79	10.44
measure_risk		*61.4*	*5.69*	167.35	119.11	88.63	31.4
Q-anonymity		–	–	*17.05*	*3.53*	10.82	9.93
GROSS.PROFIT	l-diversity	9.91	11.33	*16.76*	*4.18*	10.8	8.36
	t-closeness	0.0263	0.0256	0.001	0.0008	*0.0008*	*0.0004*
NET.PROFIT	l-diversity	10	12.6	*17.85*	*4.35*	10.71	9.62
	t-closeness	0.027	0.0264	0.0012	0.0008	*0.0008*	*0.0003*

closeness combined with Gustafson-Kessel (p-value $7.29 \cdot 10^{-5}$).

Tables 6 shows the averages of the values of each parameter ($\overline{X}$) and the distance to the optimum value (D) (when the method is not the optimum). Note that fuzzy methods perform better because their behaviour with regard to all the metrics is more stable, that means, when the method is not the best, the distance to the optimum is low.

4.2.2. Tarragona database

Tarragona dataset represents financial information of companies from Tarragona (Spain) obtained in the year 1995. It consists of 834 examples characterized by 13 attributes (*fixed.assets, current.assets, treasury, uncommitted.funds, paid.up.capital, short.term.-debt, sales, labor.costs, depreciation, operating.profit, financial.outcome, gross.profit* and *net.profit*).

In this experimentation, the performance of the attributes *gross.profit* (gross profit) and *net.profit* (net profit) as sensitive variable is studied. In the same way as in the other database, two different experiments are considered, one with each sensitive attribute.

Table 7 shows the number of experiments with the highest level of protection according to each metric and clustering method. In this table, Gustafson-Kessel is the method which performs the best with regard to k-anonymity, Q-anonymity and l-diversity. Furthermore, t-closeness has its best performance for *gross.profit* with Gustafson-Kessel method, while for *net.profit*, fuzzy c-means method gets the best results. Again we have compared the results from an statistical point of view. Thus, in this case, the proportion of times that Q-anonymity combined with Gustafson-Kessel is the best method is significantly greater than the same proportion for other methods (p-value of the proportion test $5.10 \cdot 10^{-5}$) and the same happens for l-diversity for both attributes (p-value 0.011 for *gross.profit* and 0.007 for *net.profit*). In the case of t-closeness, it is the best method when we combine it with fuzzy c-means for the attribute *net.profit* (p-value 0.026).

Table 8 shows the same values than Table 6 for Tarragona database instead. In this experimentation, the results are clearly better for the fuzzy methods with respect to k-anonymity, Q-anonymity, l-diversity and t-closeness, while *measure risk* gets better results in the crisp method. Anyway, this fact is offset by the results obtained for k-anonymity, l-diversity and t-closeness, as they are much better in both fuzzy methods than the crisp one.

5. Conclusions

In this work the goodness of protecting sensitive information by using fuzzy partitions is tested. To do that the basic metrics associated to privacy preservation are redefined in terms of fuzzy sets. The goodness of fuzzy partitions in terms of the k and Q-anonymity, l-diversity, t-closeness are checked using two standard data sets in the privacy framework when it is masked according to well-known fuzzy clustering algorithms. In addition, the risk of re-identification of a protected table masked according to different partitioning methods is also studied.

The results are promising, being possible to conclude that the use of fuzzy partitions presents an efficient alternative as masking method in the data protection framework.

However, some open problems are still presented. In fact, it is necessary to study more in depth the shape and number of fuzzy sets which better mask the released data.

Acknowledgments

Authors acknowledge financial support Grant TEC2012-38142-C04-04 from Ministry of Education and Science, Government of Spain and Grant UNOV-13-EMERG-GIJON-10 from University of Oviedo.

References

[1] H. Adeli and S.L. Hung, Machine learning – neuronal networks, genetic algorithms and fuzzy systems, John Wiley and Sons, New York (1995).

[2] R. Alcalá, J. Casillas, O. Cordón and F. Herrera, Linguistic modeling with weighted double-consequent fuzzy rules based on cooperative coevolutionary learning, *Integr Comput-Aid E* **10**(4) (2003), 343–355.

[3] A. Asuncion and D.J. Newman, UCI Machine learning repository, *School of Information and Computer Science, Univ. of California*, 2007 Available at http://www.mathworks.es/es/help/stats/kmeans.html.

[4] B. Balasko, J. Abonyi and B. Feil, Fuzzy clustering and data analysis toolbox (for use with matlab), Available at http://www.sunfinedata.com/wp-content/uploads/2009/10/FuzzyClusteringToolbox.pdf.

[5] Y. Boutalis, M. Christodoulou and D. Theodoridis, Indirect adaptive control of nonlinear systems based on bilinear neuro-fuzzy approximation, *International Journal of Neural Systems* **23**(5) (2013), 1350022 (18 pages).

[6] R. Chen, B.C.M. Fung, N. Mohammed, B.C. Desai and K. Wang, Privacy-preserving trajectory data publishing by local suppression, *Inform Sci* **231** (2013), 83–97.

[7] V. Ciriani, S. De Capitani di Vimercati, S. Foresti and P. Samarati, Microdata protection, in: *Secure Data Management in Decentralized Systems*, Y. Ting and J. Sushil, eds, Springer, 2007, pp. 291–321.

[8] I. Couso, S. Montes and P. Gil, The necessity of the strong α-cuts of a fuzzy set, *Internat J Uncertain Fuzziness Knowledge-Based Systems* **9**(2) (2001), 249–262.

[9] M. Dell'Orco and M. Mellano, A new user-oriented index based on a fuzzy inference system for quality evaluation of rural roads, *Computer-Aided Civil and Infrastructure Engineering* **28**(8) (2013), 635–647.

[10] I. Díaz, J. Ranilla, L.J. Rodríguez-Muñiz and L. Troiano, Identifying the risk of attribute disclosure by mining fuzzy rules, in: *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Methods*, E. Hüllermeier, R. Kruse and F. Hoffmann, eds, Springer, 2010, pp. 455–464.

[11] I. Díaz, L.J. Rodríguez-Muñiz and L. Troiano, Fuzzy sets in data protection: strategies and cardinalities, *Log J IGPL* **20**(4) (2011), 657–666.

[12] J. Domingo-Ferrer and U. González-Nicolás, Rational behavior in peer-to-peer profile obfuscation for anonymous keyword search, *Inform Sci* **185**(1) (2012), 191–204.

[13] C. Farkas, A. Brodsky and S. Jajodia, Unauthorized inferences in semistructured databases, *Inform Sci* **176**(22) (2006), 3269–3299.

[14] A.J. Fougeres and E. Ostrosi, Fuzzy agent-based approach for consensual design synthesis in product configuration, *Integrated Computer-Aided Engineering* **20**(3) (2013), 259–274.

[15] M. Grabisch, J.L. Marichal, R. Mesiar and E. Pap, *Aggregation Functions*, Cambridge University Press, Cambridge, 2009.

[16] W. Graf, S. Freitag, J.U. Sickert and M. Kaliske, Structural analysis with fuzzy data and neural network-based material description, *Computer-Aided Civil and Infrastructure Engineering* **27**(9) (2012), 640–654.

[17] F. Hoppner, R. Klawonn and T. Runkler, *Fuzzy Cluster Analysis*, John Wiley and Sons, 1999.

[18] F.Y. Hsiao, S.S. Wang, W.C. Wang, C.P. Wen and W.D. Yu, Neuro-fuzzy cost estimation model enhanced by fast messy genetic algorithms for semiconductor hookup construction, *Computer-Aided Civil and Infrastructure Engineering* **27**(10) (2012), 764–781.

[19] A. Hundepool, A. van de Wetering and R. Ramaswamy, μ ARGUS. Version 4.2, Available at http://neon.vb.cbs.nl/casc/Software/MuManual4.2.pdf.

[20] X. Jiang and H. Adeli, Fuzzy clustering approach for accurate embedding dimension identification in chaotic time series, *Integrated Computer-Aided Engineering* **10**(3) (2003), 287–302.

[21] E.P. Klement, R. Measiar and E. Pap, Triangular norms, Kluwer Academic Publishers, Dordrecht, 2000.

[22] V.S. Kodogiannis, M. Amina and I. Petrounias, A clustering-based fuzzy-wavelet neural network model for short-term load forecasting, *International Journal of Neural Systems* **23**(5) (2013), 1350024 (19 pages).

[23] H. Li, W. Yi and X. Yuan, Fuzzy-valued intensity measures for near-fault ground motions, *Computer-Aided Civil and Infrastructure Engineering* **28**(10) (2013), 780–795.

[24] N. Li, T. Li and S. Venkatasubramanian, T-closeness: Privacy beyond k-anonymity and l-diversity, in: *Proceedings of the IEEE ICDE* (2007), 106–115.

[25] T. Li and N. Li, Towards optimal k-anonymization, *Data Knowl Eng* **65**(1) (2008), 22–39.

[26] F. Liu and M.J. Er, A novel efficient learning algorithm for self-generating fuzzy neural network with applications, *International Journal of Neural Systems* **28**(1) (2012), 21–35.

[27] A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkitasubramaniam, l-diversity: Privacy beyond k-anonymity, *ACM Trans Knowl Discov Data* **1**(1) (2007), Article 3.

[28] N. Matatov, L. Rokach and O. Maimon, Privacy-preserving data mining: A feature set partitioning approach, *Inform Sci* **180**(14) (2010), 2696–2720.

[29] Matlab documentation center, k-means clustering, Available at http://www.mathworks.es/es/help/stats/kmeans.html.

[30] S. Montes, Fuzzy δ-ϵ-partitions, *Inform Sci* **152** (2003), 267–285.

[31] A. Ralescu, A note on rule representation in expert systems, *Inform Sci* **38**(2) (1986), 193–203.

[32] G.G. Rigatos, Adaptative fuzzy control for differentially flat MIMO nonlinear dynamical systems, *Integrated Computer-Aided Engineering* **20**(2) (2013), 111–126.

[33] S. Rokni and A.R. Fayek, A multi-criteria optimization framework for industrial shop scheduling using fuzzy set theory, *Integr Comput-Aid E* **17**(3) (2010), 175–196.

[34] E. Ruspini, A new approach to clustering, *Inform Contr* 15 (1969), 22–32.

[35] P. Samarati, Protecting respondents' identities in microdata release, *IEEE Trans on Knowl and Data Eng* **13**(6) (2001), 1010–1027.

[36] D. Shah and S. Zhong, Two methods for privacy preserving data mining with malicious participants, *Inform Sci* **177**(23) (2007), 5468–5483.

[37] E. Shmueli, T. Tassa, R. Wasserstein, B. Shapira and L. Rokach, Limiting disclosure of sensitive data in sequential releases of databases, *Inform Sci* **191** (2012), 98–127.

[38] N. Siddique and H. Adeli, Computational intelligence – synergies of fuzzy logic, neuronal networks and evolutionary computing, *Wiley, West Sussex, United Kingdom* (2013).

[39] K. Subramanian and S. Suresh, Human action recognition using meta-cognitive neuro-fuzzy inference system, *International Journal of Neural Systems* **22**(6) (2012), 1250028-15.

[40] L. Sweeney, Achieving k-anonymity privacy protection using generalization and suppression, *Internat J. Uncertain Fuzzi-*

ness Knowledge-Based Systems **10**(5) (2002), 571–588.
[41] M. Templ, A. Kowarik and B. Meindl, Statistical Disclosure Control methods for the generation of public- and scientific-use files, Available at http://cran.r-project.org/web/packages/sdcMicro/sdcMicro.pdf.
[42] D. Theodoridis, Y. Boutalis and M. Christodoulou, Dynamical recurrent neuro-fuzzy identification schemes employing switching parameter hopping, *International Journal of Neural Systems* **22**(2) (2012), 1250004-16.
[43] L. Troiano, L.J. Rodríguez-Muñiz, J. Ranilla and I. Díaz, Interpretability of fuzzy association rules as means of discovering threats to privacy, *Int J Comput Math* **89**(3) (2012), 325–333.
[44] E.I. Vlahogianni and M.G. Karlaftis, Fuzzy-entropy neural network freeway incident duration modeling with single and competing uncertainties, *Computer-Aided Civil and Infrastructure Engineering* **28**(6) (2013), 420–433.
[45] L. Yan and Z.M. Ma, Conceptual design of object-oriented databases for fuzzy engineering information modeling, *Integrated Computer-Aided Engineering* **20**(2) (2013), 183–197.
[46] L. Yan and Z.M. Ma, Extending engineering data model for web-based fuzzy information modelingl, *Integrated Computer-Aided Engineering* **20**(4) (2013), 407–420.
[47] Y. Zhang and H. Ge, Freeway travel time prediction using takagi-sugeno-kang fuzzy neural network, *Computer-Aided Civil and Infrastructure Engineering* **28**(8) (2013), 594–603.
[48] S. Zhong, Privacy-preserving algorithms for distributed mining of frequent itemsets, *Inform Sci* **177**(2) (2007), 490–503.