What Could Fuzzy Logic Bring to Statistical Information Systems?

Miroslav Hudec[a] | *INFOSTAT, Bratislava*

Abstract

The aim of the paper is to present the applicability of the fuzzy logic for statistical information systems in order to improve work with statistical data. The improvement offers the approximate reasoning in order to solve problems in a way that more resembles human logic. The paper examines the fuzzy logic approach, emphasizes situations when the two-valued (crisp) logic is not adequate and offers solutions based on fuzzy logic. The first step of using data is its selection from a database. Although the Structured Query Language (SQL) is a very powerful tool, it is unable to satisfy needs for data selection based on linguistic expressions and degrees of truth. For this purpose the fuzzy generalised logical condition (GLC) was developed. Later researches have shown that the GLC formula is suitable for other processes concerning data, namely data classification and data dissemination.

Key words:

fuzzy database query,
fuzzy generalised logical condition,
applicability of fuzzy query

INTRODUCTION

The following statement holds for many cases: If something is precisely defined it is easy to implement but it could be far from reality; if something contains ambiguities and uncertainties in its definition it is harder to implement but it is better connected to reality. This statement is generally true but there are methodologies capable to easier implement ambiguities and uncertainties.

One of these methodologies is fuzzy set and fuzzy logic. "Fuzzy logic allows us to bring the operation of information systems closer to the working methods of humans." [6]. Users frequently deal with terms such as high unemployment rate, the majority of, low migration level, etc. These terms include a certain vagueness or uncertainty that information systems based on two-valued logic {true, false} do not understand and therefore cannot use. The fuzzy set theory works with the gradation, uncertainty and ambiguity described by linguistic expressions when sharply defined criteria could not be created. More about this kind of vagueness and uncertainty can be found in [24].

Statistics is a promising area to develop and implement the fuzzy logic concept. Large data collections are mainly stored in relational databases. Users need this data for the analysis, decision making or just to find interesting information. In many cases the rigid or precise definition of an analysed task cannot be created or the user wants to obtain additional information.

The first step of using data is in many cases its selection from relational databases. The SQL is used for this purpose. Although the SQL is a powerful tool, it uses the two-valued logic in the selection process. It causes that the small er-

[a] *INFOSTAT, Dubravska cesta 3, 845 24 Bratislava, Slovakia, e-mail: hudec@infostat.sk*

ror in data values or in cases when the user cannot unambiguously define the criterion by crisp boundaries may involve some inadequately selected or non-selected data. As a solution, the use of fuzzy logic is proposed and described in the paper. This area of research is not new, but there are still many possibilities for the improvement of existing approaches and for creating new approaches. Some fuzzy query implementations have been designed e.g. [2] and [22]. Although there are some variations according to the particularities of different implementations, the answer to a fuzzy query sentence is generally a list of records, ranked by the degree of matching [3]. Issues and perspectives of fuzzy querying can be found in [14]. In order to bring benefit of the fuzzy logic to database users and to make an easy to use querying tool the generalized logical condition (GLC) for database queries was created in [9]. The implementation of the GLC for statistical databases was discussed in [10].

Later researches have shown that the GLC formula can be used for other processes concerning data. In our researches we are mainly interested in information systems improvements for data classification and data dissemination areas. Application for municipalities classification by fuzzy expert systems in the Slovak Republic was proposed in [12]. Another way of classification has been found during work on fuzzy database queries. The GLC could be used for data classification by generating queries from fuzzy rules [11]. Generating queries from fuzzy rules is an active research field. Some fuzzy classification implementations have been proposed in [3] and [21]. The last two papers contain information about other researches and implementations in this field.

The data dissemination becomes a very important area especially for organisations that provide wide variety of data to professionals and broad audience. One of dissemination channels is the internet. Many articles and books about website design and appropriate design of tables and graphs can be found, for example in [19] and [5] respectively. In our paper advantages of the fuzzy approach in the process of finding and selecting adequate data for statistical websites are pointed out.

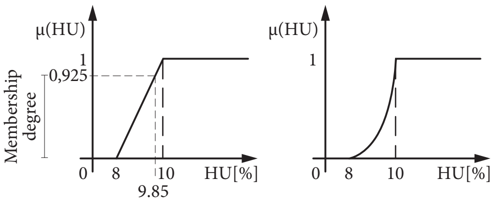
The paper is organised as follows: The crucial differences between crisp and fuzzy logic are reviewed in Section 2. SQL and fuzzy database queries are examined in section 3. The disadvantage of SQL is pointed out and an alternative approach based on the fuzzy GLC is analysed. In the last part of this section the proposed concept is demonstrated by a case study. Section 4 emphasises that developed GLC and fuzzy queries can be used for many other purposes. The usefulness of the GLC for data classification and dissemination is discussed in this section. Finally some conclusions are drawn.

1 FUZZY LOGIC

The core of both crisp logic and fuzzy logic is the idea of a set. In the crisp set theory an element belongs or does not belong to a set. For example, consider a set called high unemployment (HU) defined as follows: HU = {x| unemployment(x) ≥ 10%} where x is a region. It means that region with 9.9% unemployment does not belong to the HU but region with 10% belongs. These constraints are drawback when the boundaries between values of some attributes are continuous.

The concept of fuzzy sets was initially introduced in [23] where was observed that more often than not, precisely defined criteria of belonging to a set could not be defined. The fuzzy set and logic theory brings a paradigm in work with the gradation, uncertainty and ambiguity described by linguistic expressions. This gradation is described by a membership function μ valued in the interval [0, 1]. The HU example can be presented by fuzzy sets shown in Figure 1. User could define that the unemployment equal and higher than 10% is HU, the unemployment smaller than 8% definitely is not HU and unemployment between 8% and

Figure 1 Fuzzy sets for high unemployment concept



Source: own research

10% partially belongs to the HU concept. The closer is the unemployment to 10% the stronger it belongs to the high unemployment concept. The fuzzy set gives answer for the following question: How compatible is 9.85% unemployment with the HU concept? The answer is 0.925 or territorial unit with 9.85% unemployment rate is a very strong member of the HU concept. If territorial unit's unemployment is 9%, it is a moderate member of the HU concept.

2 DATABASE QUERIES

2.1 SQL and its limitation

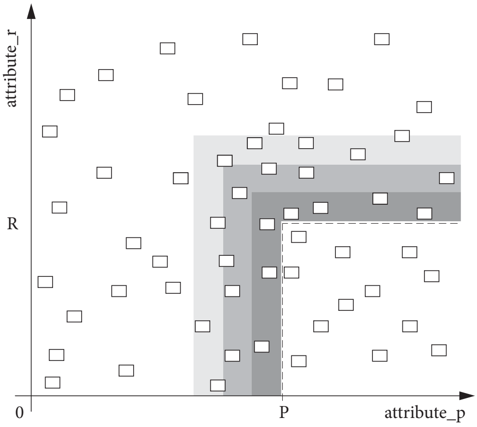
The SQL was initially developed in [4]. Since then the SQL has been used in many relational databases and information systems for data selection. The use of SQL may be regarded as one of the mayor reasons for the success of relational databases in the commercial world [20].

Generally speaking, users search databases in order to obtain data needed for analysis, decision making or just to satisfy their curiosity. Situations when constraints of crisp logic in querying processes may occur are examined by the following example:

```
select <attribute(s) list>
from <table(s) list>
where attribute_p > P and attribute_r < R;
```

The best way how to describe limitations of a SQL query is the graphic mode shown in Figure 2. Values P and R delimit the space of selected data. The user cannot obtain any information about records that are close to meet the query criterion (areas marked with grey shadows). The area marked with the darkest grey shadow contains records that almost meet the intent of the query. It means that the record would not be selected even if it is extremely close to meet the query criterion. Records belonging to shadowed areas could be potential customers and direct marketing could attract them or territorial units which almost satisfy criterion for some financial support for example. In case of no data is selected by SQL, there is not any information concerning possible records that almost meet the query criterion. This is the penalty paid to use the crisp logic in selection process.

Figure 2 The result of the SQL query



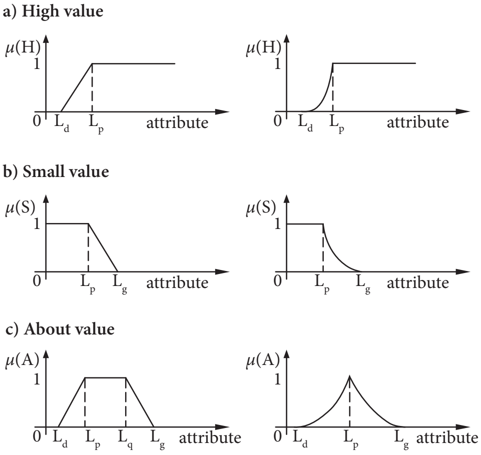
Source: own research

The new way of evaluating the *WHERE* clause of a SQL query and its further applicabilities are explained in next subsections.

2.2 Fuzzy query idea

SQL conditions in queries contain these comparison operators: > , < , = , ≠ and *between* when numerical attributes are used. These crisp logical comparison operators are adapted for fuzzy queries in the following way: operator > (greater than) was improved with fuzzy set "High value" (Figure 3a), operator < (less than) was improved with fuzzy set "Small value"

Figure 3 Fuzzy sets



Source: own research

(Figure 3b) and operator = (equal) was improved with fuzzy set "About value" (Figure 3c). Operator ≠ is the negation of the operator = so this operator is not further analysed. Analogous statement is valid for the operator *between* because it is similar to the operator = from the fuzzy point of view.

For the further reading it is important to define the Query Compatibility Index (QCI). The QCI is used to indicate how the selected record satisfies a query criterion. The QCI has values from the [0, 1] interval with the following meaning: 0 – record does not satisfy the query, 1 – record fully satisfies the query, interval (0, 1) – record partially satisfies the query with a distance to the full query satisfaction.

According to the above mentioned facts, the example of a fuzzy query e.g. to find appropriate areas for tourism has the following form:

select district
from table
where air_pollution is Small and number_of_
sunny_days is High;

The meaning of a fuzzy query is obvious at first glance because it is expressed with linguistic expressions. The shape of membership function (μ(x)) (Figure 3) can be adjusted according to user's requirements without changing the meaning of a query. In this example the fuzzy set with $L_p = 10$ and $L_g = 15$ units of measured pollutant and shape as from Figure 3b) on the left side describes the small air pollution concept. The fuzzy set with $L_d = 140$ and $L_p = 150$ days and shape as from Figure 3a) on the left side describes the high number of days with sunshine concept. The result of this query is in the table 1.

In this example μ(P) denotes the membership degree to the small pollution fuzzy set and μ(S) denotes the membership degree to the high number of days with sunshine fuzzy set. The QCI (calculated as a minimum) represents membership degree to the small air pollution and high number of sunny days concept. If user wants to make some activity in the appropriate district and it is not possible to realise it in districts D7, user can choose the district D5 that almost satisfies the intent of the query. Again, if it is not possible to choose the district D5 the next choice is D2 and so on. It is important to emphasize that ranking is not done by one indicator, their linear combination by weighted coefficients, etc. Both indicators have the same importance and ranking is done according to the satisfying the concept created in the query criterion.

2.3 Fuzzy query realisation

The starting point of our research was the following premise: To make easy to use data selection by concepts and to access to relational databases in the same way as SQL does. Suggested querying process consists of two main steps. In the first step all records that have the membership degree to the condition defined by linguistic expressions in the *WHERE* clause greater than zero (QCI > 0) are selected from database. For this purpose the GLC for the *WHERE* part of the SQL was created and described in [9]. The GLC has the following structure:

$$\text{WHERE} \bigotimes_{i=1}^{n} (a_i \circ L_{xi}) \qquad (1)$$

where *n* denotes number of attributes with fuzzy constraints in a *WHERE* clause of a query,

$$\otimes = \begin{cases} and \\ or \end{cases}$$

where *and* and *or* are fuzzy logical operators, and

Table 1 Areas conductive to the tourism

District	Pollution (P)	μ (P)	Number of sunny days (S)	μ (S)	QCI	
D2	9.5	1	147	0.7	0.7	←
D3	11	0.8	145	0.5	0.5	
D5	10.2	0.96	149	0.9	0.9	←
D7	8.2	1	161	1	1	
D8	14.1	0.18	160	1	0.18	

Source: own research

$$a_i \circ L_{ix} = \begin{cases} a_i > L_{di}, & \text{a}_\text{i} \text{ is High} \\ a_i < L_{gi}, & \text{a}_\text{i} \text{ is Small} \cdot \\ a_i > L_{di} \text{ and } a_i < L_{gi}, & a_i \text{ is About} \end{cases}$$

where a_i is a database attribute, L_d is the lower bound and L_g is upper bound of a linguistic expression described by fuzzy set. Two types of fuzzy set for High, Small and About expressions are shown in Figure 3.

In this step lower and/or upper bounds of linguistic expressions (fuzzy sets) are used as parameters for database query criteria. Let take the *WHERE* clause from the previous query:

> where air pollution is Small and number of sunny days is High.

Parameters L_p and L_g are used to define meaning of the subcriterion "air pollution is Small". User could state that district with measured pollutant less than 10 units fully belongs to the analysed concept and the parameter L_p set this state: $L_p = 10$ units. District with air pollution between 10 and 15 partially belong to the concept air pollution is Small. The closer is the air pollution to 10 units the stronger it belongs to the small air pollution concept. User could state that district with air pollution higher than 15 units does not belong to small air pollution concept and the parameter L_g is used to set this state: $L_g = 15$ units. Similar discussion holds for the subcriterion "number of sunny days is High".

According to the parameters of fuzzy sets and the GLC (1), fuzzy query is converted into the following SQL structure:

> where air pollution < 15 and number of sunny days > 140.

This *WHERE* clause ensure that query selects all records with QCI > 0 from a database.

In the second step the chosen analytical form of the fuzzy set is used to calculate the membership degree of each selected record to appropriate fuzzy set e.g. pollution value to concept of low pollution and number of sunny days to concept of high number of sunny days. Finally, the QCI value for each selected record is calculated. When a classical query contains more than one condition in the *WHERE* clause *and* and *or* logical operators are used. In classical case there exists only one logical function for *and* and *or* operators because the subcriterion is satisfied (value 1) or not (value 0). In fuzzy logic there exist many functions describing *and* operator (these functions are called t-norms) and *or* operator (these functions are called t-conorms) because each of subcriterions can be fully or partially satisfied. More about t-norm and t-conorm function could be found in e.g. [15]. For example the territorial unit satisfies the high number of days with sunshine concept with 0.5 and the low air pollution with 0.8. Both conditions are partially satisfied so the {0, 1} logic is not useful. It is needed to combine membership degrees so that the total result of a query can be expressed. The following t-norm functions can be used [18] for logical *and* operator:

- minimum:

$$\text{QCI} = \min(\mu_\text{i}(\text{a}_\text{i})), \quad \text{i} = 1,\ldots, \text{n} \tag{2}$$

- product:

$$\text{QCI} = \prod_{\text{i}=1}^{\text{n}} (\mu_i(\text{a}_\text{i})), \tag{3}$$

- bounded difference (BD)

$$\text{QCI} = \max(0, \sum_{i=1}^{n} \mu_i(a_i) - n + 1) \tag{4}$$

The following t-conorm functions can be used [18] for logical or operator:

- max

$$\text{QCI} = \max(\mu_i(\text{a}_\text{i})), \quad \text{i} = 1,\ldots, \text{n} \tag{5}$$

- bounded sum (BS)

$$\text{QCI} = \min(1, \sum_{i=1}^{n} \mu_i(a_i)) \tag{6}$$

where $\mu_i(a_i)$ denotes the membership degree of the attribute a_i to the *i*-th fuzzy set. The min t-norm takes into account the lowest value of membership degrees to fuzzy sets (0.5 in previous example). The product t-norm takes into account all membership degrees and balances the query truth membership value across each of conditions in the *WHERE* clause (0.4 in previous example). The whole process concerning data selection by the GLC can be found in [9].

2.4 Case study

Data from the Urban and Municipal Statistical database [1] are used for case study. This database is in official use at the Statistical Office of the Slovak Republic. In this case study, districts with high length of road and small area size are sought. The high road infrastructure density is analysed as an illustrative example. The query has the following form:

```
select district
from table
where roads is High and area is Small;
```

The road length indicator is represented by High fuzzy set with these parameters $L_d = 150$ km and

Table 2 Result of fuzzy query

District	Roads [km]	Area [km^2]	µ (Road)	µ (Area)	QCI
Bratislava I	335.1	9.6	1	1	1
Senec	269.1	359.9	1	1	1
Piešťany	305.6	381.1	1	1	1
Myjava	563.9	327.4	1	1	1
Púchov	320.9	375.4	1	1	1
Bytča	231	281.6	1	1	1
Kysucké Nové Mesto	269.9	173.7	1	1	1
Detva	567.2	449.2	1	1	1
Žarnovica	366.6	425.5	1	1	1
Považská Bystrica	**324.5**	**463**	**1**	**0.913**	**0.913**
Sabinov	**220.8**	**483.5**	**0.888**	**0.777**	**0.777**
Šaľa	**206.9**	**355.9**	**0.713**	**1**	**0.713**
Poltár	**207.4**	**476.1**	**0.713**	**0.826**	**0.713**
Ilava	**205.8**	**358.5**	**0.7**	**1**	**0.7**
Dolný Kubín	**197.8**	**491.8**	**0.6**	**0.721**	**0.6**
Žiar nad Hronom	*249.8*	*517.6*	*1*	*0.549*	*0.549*
Zlaté Moravce	*226.4*	*521.2*	*0.95*	*0.525*	*0.525*
Hlohovec	*187.1*	*267.2*	*0.463*	*1*	*0.463*
Pezinok	*176.9*	*375.5*	*0.338*	*1*	*0.338*
Bánovce nad Bebravou	*172.5*	*461.9*	*0.275*	*0.921*	*0.275*
Partizánske	*168*	*301.2*	*0.225*	*1*	*0.225*
Tvrdošín	*164.9*	*478.9*	*0.188*	*0.807*	*0.188*
Svidník	*164.5*	*549.6*	*0.175*	*0.336*	*0.175*
Nové Mesto nad Váhom	*528.5*	*580*	*1*	*0.133*	*0.133*
Gelnica	*163.6*	*584.4*	*0.175*	*0.104*	*0.104*
Krupina	*334.9*	*584.9*	*1*	*0.101*	*0.101*
Levoča	*157.1*	*357.2*	*0.088*	*1*	*0.088*
Spišská Nová Ves	*388.9*	*587.4*	*1*	*0.084*	*0.084*
Topoľčany	*371.8*	*597.7*	*1*	*0.015*	*0.015*

Source: [1]

$L_p = 230$ km and the shape as from Figure 3a) on the left side. The Small fuzzy set with parameters $L_p = 450$ km^2 and $L_g = 600$ km^2 and shape as from Figure 3b) on the left side describes the district area indicator.

The result of fuzzy query is shown in Table 2. The value of min t-norm (2) is used for the calculation of the QCI. The Table 2 shows nine districts fully satisfying the query; one district is extremely close to satisfy the query (marked with the stronger bold text) and another five districts are close to meet the query criterion. These five records are marked with the lighter bold text. It means for example that even small changes in attributes could imply that another district fully satisfies the query. If SQL were used, this additional valuable information would remain hidden. Territorial units are distinguished according to gradation of belonging to the concept (the query criterion).

If SQL were used, the criterion would be as follows:

where roads > 230 and area < 450.

The result of this criterion is shown in Table 3. The difference between information in the Table 2 and Table 3 is obvious. Records marked as bold and italic in Table 2 are not selected by the SQL query and the last three rows from the Table 2 is not possible to calculate because SQL query selects data only whereas fuzzy query selects data and calculates additional information.

Fuzzy queries reduce the risk of obtaining empty answer. In situation when no data is selected by the SQL, fuzzy query can inform that there are some records that almost meet the query criterion. It means that all data marked with bold and italic text in Table 2 are selected only and the distance to full query satisfaction for these records is calculated. It is not need to rearrange the fuzzy query in order to select some records.

2.5 Some fuzzy query characteristics

The SQL is a very powerful and useful query language, but only a query language. In this research the core of SQL remains intact and the extension is done to improve the querying process. Adding some flexibility to the SQL increases effectiveness and comprehensibility of the data selection. As metadata are used to explain the meaning of figures, linguistic expressions are used to explain the meaning of a query. The fuzzy approach improves the SQL with approximate reasoning. The intent of query based on fuzzy logic is not to select more data but to select better data. The advantages of this approach for users are as follows [10]:

Table 3 Result of SQL query

District	Roads [km]	Area [km^2]
Bratislava I	335.1	9.6
Senec	269.1	359.9
Piešťany	305.6	381.1
Myjava	563.9	327.4
Púchov	320.9	375.4
Bytča	231	281.6
Kysucké Nové Mesto	269.9	173.7
Detva	567.2	449.2
Žarnovica	366.6	425.5

Source: [1]

- the connection to the database and the data accessing do not have to be modified,
- users do not need to learn a new query language,
- the querying process supports the (quasi) natural language,
- presenting of obtained data is in similar way as from SQL but with additional valuable information,
- users see data "behind the corner" (grey areas on Figure 2 and bold text in Table 2).

Database querying languages based on the fuzzy logic need additional calculations in comparison with SQL counterpart. This constatation also holds for the methodology suggested in this article. The first additional step consists of conversion from linguistic expressions to the SQL structure. The second additional step, activated immediately after records are selected from database, consists of QCI calculation for each of selected record. This additional amount of calculation is balanced with additional information obtained from databases.

There is no competition between querying based on crisp and fuzzy logic. A fuzzy database query provides flexibility for the inclusion of records that are close to meet the query criterion (potential candidates) and to calculate additional valuable information. SQL database queries are useful when clean and exact boundary between selected and non selected data is required.

3 FUZZY QUERIES AND THEIR FURTHER APPLICABILITY IN STATISTICS

The statement that fuzzy querying engines gives new possibilities for data selection has proven in the previous chapter. This chapter draws attention to applicability of fuzzy queries for a broader usage. In this section the improvements of data classification and data dissemination by fuzzy logic and the GLC are examined.

There exist many other applications of fuzzy logic and fuzzy queries. Some of them are:

- case based reasoning realised inside relational database using fuzzy query approach [16],
- extension of fuzzy query for data mining and knowledge discovery [17],
- fuzzy logic approach can be used in GIS for many purposes from finding locations to spatial data analysis. One example can be found in [13].

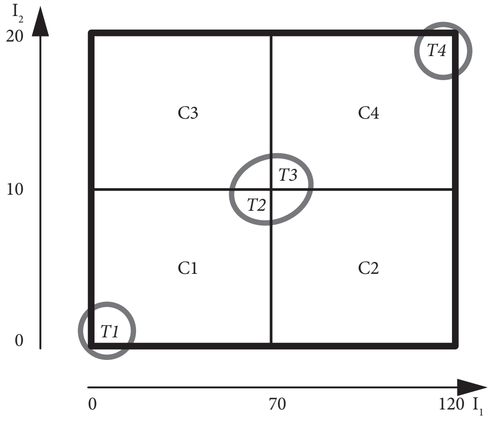
The above mentioned areas where fuzzy logic and fuzzy queries could be used are mentioned only to point out how wide are possibilities to use fuzzy logic in information systems.

3.1 Data classification

3.1.1 Crisp and fuzzy classification comparison

In classification by crisp tools, classes have sharp boundaries. If values of indicators are similar for two objects (customers, territorial units), they are similar too but it could imply that objects may fall into different classes. The classification diagram presented in Figure 4 shows this situation in graphical mode. Objects are divided into four classes from class C1 (the smallest) to class C4 (the biggest). This method treats the top rated object T4 in the same way as T3. Units T2 and T3 have similar indicators values. However, T2 and T3 are treated in different classes.

Figure 4 Crisp classes



Source: own research

Expert systems offer a good support for classification but limitations of crisp logic may occur. The following question arises: How to solve this problem without additional calculation from user's point of view? The answer is fuzzy logic. In fuzzy classification classes do not have sharp boundaries and a classified object can belong to more than one overlapping class. Belonging to a fuzzy class depends of the membership degree to the relevant class.

The fuzzy approach gives two main ways for solving classification tasks: fuzzy systems and generating fuzzy queries from previously created fuzzy rules. The first way is an extension of expert systems by fuzzy sets and fuzzy logic. Fuzzy systems and their applicability are examined in details in [18]. The fuzzy inference system (FIS) from the Mat Lab software was used to create and solve municipalities classification model [7] and [12].

By reason that the emphasis of this paper is on the GLC and its applicability, classifications by fuzzy systems are not further considered in the paper. The idea for classification by the GLC has been found during work on fuzzy database queries. Researches have shown that the GLC formula (1) could be used in data classification [11]. Queries are equivalent with the *IF* part of the rules and result of the query are records that fully or partially belong to the output class representing the *THEN* part of the rules. Classification by the fuzzy system and by the GLC has the same rule base structure but ways how these rules are calculated in order to obtain solution is different.

3.1.2 Classification by the GLC

"Fuzzy queries sentences are structured definitions of fuzzy concept. Under this assumption, fuzzy queries can be automatically generated by fuzzy rule based classifiers" [3]. This paper illustrates the classification using above described fuzzy queries and the GLC. The difference is in the added clause *CLASSIFY_INTO*. The *CLASSIFY_INTO* clause specifies the name of the output class to which selected records satisfying query are classified. This membership degree is also membership degree to the appropriate output class. The structure of a query is as follows:

classify_into [class$_c$]
select <attribute(s) list>
from <table(s) list>
where $\left[\bigoplus_{j=1}^{m}\bigotimes_{i=1}^{n}(a_i \circ L_{xij})\right]_C$

where the logical operator ⊗ from (1) describes *IF* part of the rule, ⊕ is the logical *or* operator that merges those *m IF* parts of rules that have the common *THEN* part or the same output class *c*, *n* is the number of attributes inside the *IF* part of the rule.

Object can belong to more than one class with different membership degrees. The rank of object is calculated by the aggregation of class coefficient where object belongs and its membership degree to these classes respectively using the following equation:

$$R_O = \sum_{l=1}^{L} \mu_{Ocl} K_l \tag{7}$$

where L is number of classes, μ_{Ocl} is the membership degree of object O to class C_l and K_l is the parameter describing class C_l.

Advantages of this approach are as follows [10]:

- queries select only those records that will be classified. Records that do not belong to any class are not needlessly selected;
- data preparation to the adequate input vector or matrix like for fuzzy systems is not needed;
- presentation of results in a useful and understandable form for example in the xls format could be easy implemented.

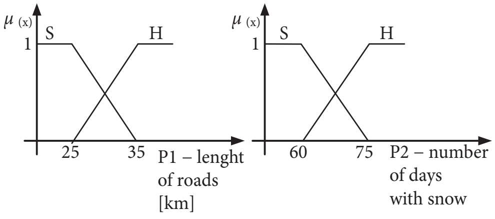
3.1.3 Case study

The classification feasibility of the GLC is illustrated with the purpose of rough planning of road maintenance requirements in winter. Detailed classification model which contains three indicators fuzzified into five fuzzy sets and 93 rules in the rule base can be found in [12] where the model was solved by fuzzy system using the Mat Lab software. In order to present classification by fuzzy queries and the GLC on illustrative example, the model is reduced. Data from the same database as for the fuzzy query case study was used. In this example two indicators are included and fuzzified into two fuzzy sets: length of roads in kilometres (Road) and number of days with snow (Snow). These sets are shown in Figure 5.

This example contains four fuzzy rules with the following structure:

- if Road is Small and Snow is Small Then Maintenance is Small;
- if Road is Small and Snow is High Then Maintenance is Medium;
- if Road is High and Snow is Small Then Maintenance is Medium High;
- if Road is High and Snow is High Then Maintenance is High.

Figure 5 Fuzzy sets small (S) and high (H) for Roads and Snow indicators



Source: own research

According to the rule base and the GLC (1) four fuzzy queries are created. The query for the Small output class has the following form:

```
classify_into Small
select municipality
from table
where roads is Small and snow is Small;
```

The percentage of requirements parameter (K) for the winter road maintenance can be associated with each fuzzy output class: for instance class S (Small) gets 10%, class M (medium) gets 35%, class MH (Medium High) gets 65% and municipalities from class H (high) gets 90% from considered amount of resources. Table 4 shows ranking results for some municipalities.

The fuzzy classification allows softer classification and ranking among municipalities (municipalities marked with the bold text in Table 4). In case of crisp classification these municipalities may be classified into different classes and it will cause difference between required needs and obtained resources. To avoid this disadvantage the user has to create very high number of output classes and rules if he wants to use crisp classification tools.

Territorial units that partially belong to more than one class are treated in all classes where they have partially membership. If data values of attributes are similar for territorial units, they are similar treated and get nearly the same percentage of resources for example.

The similar reason holds for choosing between fuzzy and crisp classification and between fuzzy and crisp selection. Fuzzy classification provides flexibility for the classification with the gradation of belonging to overlapped classes. Crisp classification is useful when clean and exact boundary between output classes is required.

Although mathematics based on fuzzy sets has greater expressive power than classical mathematics based on crisp sets, the usefulness depends critically on our capability to construct appropriate number of fuzzy sets, describe their membership functions and create all relevant fuzzy rules.

3.2 Data dissemination

Dissemination of statistical data targeting web-based audience is one of important tasks of statistical organisations. This poses a significant challenge to the statistical organisation to provide the suitable website design, accuracy, timeliness, and reliability of data and metadata. Metadata (especially descriptive metadata and metadata assisting in the navigation and search) are important elements for data dissemination. The metadata facilitate legibility and apprehensibility of the disseminated data and ensure the correct interpretation of presented data.

These metadata are also used for creation of database queries, more precisely in the projection (which columns from tables are included in query)

Table 4 Result of fuzzy classification

Municipality	Coefficient of needs (R)
Banská Bystrica	0.9
Banská Štiavnica	0.9
Zvolen	0.9
Detva	0.9
Donovaly	**0.845**
Lučenec	**0.75**
Cerovo	**0.68**
Fiľakovo	0.65
Rimavská Sobota	0.65
Horný Tisovník	**0.515**
Jasenie	0.35
Banský Studenec	0.35
Kremnica	0.35
Podhorie	0.35
Sliač	0.35
Skerešovo	**0.33325**
Leváre	**0.31675**
Pôtor	**0.31675**
Jelšava	**0.26675**
Vinica	**0.25**
Rapovce	**0.21675**
Bottovo	**0.16675**
Radzovce	0.1
Hostice	0.1
Nenince	0.1
Dudince	0.1

Source: own research

and the selection (which conditions have to be satisfied to extract a record from the database). In the projection phase the user chooses interesting indicators (the *SELECT* statement of a query). In the selection phase the user is limited by the properties of the crisp logic in the *WHERE* clause of a query. It means that the record either satisfies the intent of a query or do not satisfies it. This logic does not permit any other possibility. In some cases this property of the crisp logic is desirable. For example, the user wants to select all municipalities belonging to the district A. The meaning and logic of this query is two-valued (municipality fully belongs or fully does not belong to the district A).

In cases when the logic of a query cannot be limited by crisp logic, the fuzzy approach gives a solution. For example, when the user wants to find towns with good living conditions, the user can describe preferences by linguistic expressions. The output of a query is softly ranked towns according to the previously created preferences.

In [21] fuzzy classification interpreter and editor have been implemented as Java Servlets. A similar approach could be applied for fuzzy selection. The GLC was tested on desktop application. The idea of broadening this approach to websites might be very perspective and is under consideration. The interesting candidate for testing and implementing data dissemination by fuzzy queries is the population and housing censuses in Slovakia on the website [8]. The essence of fuzzy queries is reducing or eliminating the communication barrier between the human and the computer during querying process. Another reason for this development is the fact that the goal of many websites is to target broad audience. Many users of websites are not familiar with limitations of SQL and they expect data selection process to be closer to working methods of humans. Providing a query by linguistic expressions gives natural way for database queries creation and websites could become more user friendly oriented in processes of data selection.

CONCLUSION

It is proven in our research that the proposed fuzzy logic approach can improve work with statistical information systems. If crisp sets and sharp boundaries in queries are used the result may involve some inadequately selected data, e.g. in cases when the user cannot unambiguously define the criteria by crisps values. The SQL requires the crisp specification of a query criterion, while for users a query is better describable in terms of a natural (or quasi) natural language with ambiguities and uncertainties. This is one of reasons why the research has started with database querying improvements by fuzzy logic. As an output of this research, the GLC was created. In this way, queries based on linguistic expressions on client side are supported and are accessing relational databases in the same way as the SQL. No modification of databases has to be undertaken.

The goal of query based on fuzzy logic is not to select more data but to select more representative data. The fuzzy logic approach is not only more natural for users, but it is also more powerful. Data is selected according to the gradation of satisfying a query criterion. Database querying languages based on fuzzy logic demand additional calculations in comparison with SQL counterpart. This additional amount of calculation is balanced with additional information obtained from database.

The software for fuzzy selection based on the GLC has been developed on prototype level. More precisely, required items needed for case studies were realized. The stress was on research and creation of equations to describe the fuzzy logic and its potentiality.

Later researches have shown that the GLC can be used for data classification and dissemination. Fuzzy classification approach gives users the possibility to include the approximate reasoning into the classification problem by creating fuzzy *IF-THEN* rules. These fuzzy rules are converted into fuzzy queries and solved using the GLC. Data dissemination on websites is mentioned as an interesting field where queries based on the GLC could be realised.

It is important to point out that there is no competition between computing by crisp logic and computing by fuzzy logic. A fuzzy query provides flexibility when user cannot unambiguously define the criterion by crisps boundaries or user can not expressly prove why the chosen boundary value is the best one. Moreover, selection of relevant entities from data sets is more flexible, allowing examination of records that clearly meet the criteria, as well as those that almost meet the given criteria. SQL database queries are useful when a clean and exact boundary between selected and non selected data is required and user is interested in data which clearly meet given criteria only. The similar statement holds for the classification. In dissemination the nature of searched data or information predetermines the use of SQL or fuzzy query. It is on the user to decide which approach is better for the particular task. Although mathematics based on fuzzy sets and fuzzy logic has greater expressive power than classical mathematics based on crisp sets and crisp logic, the usefulness depends critically on our capability to construct appropriate membership functions of linguistic expressions and create relevant fuzzy rules for each particular task.

Metadata are used to explain the meaning of the indicator and its values. The similar constatation can be told for fuzzy data selection: linguistic expressions are used to explain the meaning of a query.

References

[1] BENČIČ, A., HUDEC, M. *MOŠ/MIS–Urban and municipal statistics project and information system of the Slovak Republic.* SYM-OP-IS, XXI-32--XXI-35, 2002.

[2] BOSC, P., PIVERT, O. SQLf Query Functionality on Top of a Regular Relational Database Management System. In: Pons M, Vila M A and Kacprzyk J (eds.). *Knowledge Management in Fuzzy Databases.* Physica Publisher, Heidelberg, 2000. pp 171–190.

[3] BRANCO, A., EVSUKOFF, A., EBECKEN, N. *Generating Fuzzy Queries from Weighted Fuzzy Classifier Rules.* ICDM workshop on Computational Intelligence in Data Mining, 2005. 21–28.

[4] CHAMBERLIN, D., BOYCE, R. SEQUEL: *A Structured English Query Language.* ACM SIGMOD Workshop on Data Description, Access and Control, 1974. 249–264.

[5] FEW, S. *Show me the numbers – Design tables and graphs to enlighten.* Oakland: Analytic Press 2004.

[6] GALINDO, J., URRUTIA, A., PIATTINI, M. *Fuzzy Databases: Modeling, Design and Implementation.* Hershey: Idea Group Publishing 2006.

[7] HUDEC, M., VUJOŠEVIĆ M. *Fuzzy systems and neuro-fuzzy systems for the municipalities classification.* Eurofuse anniversary workshop on "Fuzzy for Better", 2005. 101–110.

[8] HUDEC, M., BÜCHLER, P. *Metadata and website design for statistical data dissemination. Management*, No 52, 2009. 23–30.

[9] HUDEC, M. *An Approach to Fuzzy Database Querying, Analysis and Realisation.* Computer Science and Information Systems Vol. 6, No. 2, 2009. 124–140.

[10] HUDEC, M. *Soft computing techniques for statistical databases.* Meeting on the Management of Statistical Information Systems, 2009. <http://www.unece.org/stats/documents/ece/ces/ge.50/2009/wp.22.e.pdf>.

[11] HUDEC, M., VUJOŠEVIĆ, M. *Selection and Classification of Statistical Data Using Fuzzy Logic.* NTTS Conferences on New Techniques and Technologies for Statistics, 2009. 186–195.

[12] HUDEC, M., VUJOŠEVIĆ, M. A fuzzy system for municipalities classification. *Central European Journal of Operations Research,* Vol.18, No. 2, 2010. 171–180.

[13] IOANNIDIS, C., HAZICHRISTOS, T. *A municipality selection proposal for the expansion of the Hellenic cadastre using fuzzy logic.* Spatial information management, experience and visions for the 21st century, 2000. <http://www.fig.net/com_3_athens/>.

[14] KACPRZYK, J., PASI, G., VOJTÁŠ, P., ZADROZNY, S. Fuzzy querying: Issues and perspectives. *Kybernetika*, Vol. 36, No. 6, 2000. 605–616.

[15] KLIR G., YUAN B. *Fuzzy sets and fuzzy logic, theory and applications.* Prentice Hall: New Jersey 1995.

[16] PORTINALE, L., VERRUA, A. *Exploiting Fuzzy-SQL in Case-Based.* Florida Artificial Intelligence Research Society Conference, 2001. 103–107.

[17] RASMUSSEN, D., YAGER, R.R. Summary SQL – A Fuzzy Tool for Data Mining. *Intelligent Data Analysis,* 1, 1997. pp. 49–58.

[18] SILER, W., BUCKLEY, J. *Fuzzy expert sytems and fuzzy reasoning.* New Jersey: John Wiley & Sons, 2005.

[19] *United Nations Statistical Commision and Ecconomic Commision for Europe. Best practices in designing websites for dissemination of statistics.* Conference of European statisticians, Methodological material, Geneva, 2001.

[20] URRUTIA, A., PAVESI, L. *Extending the capabilities of database queries using fuzzy logic.* Collecter-LatAm, 2004. <http://www.collecter.org/archives/2004_October/06.pdf>.

[21] WERRO, N., MEIER, A., MEZGER, C., Schindler, G. *Concept and Implementation of a Fuzzy Classification Query Language.* International Conference on Data Mining, 2005. 208–214.

[22] WANG, T.C., LEE, H.D., CHEN, C.M. *Intelligent Queries based on Fuzzy Set Theory and SQL.* Joint Conference on Information Science, Salt Lake City, 2007. 1426–1432.

[23] ZADEH, L. Fuzzy Sets. *Information and Control,* No. 8, 1965. 338–353.

[24] ZIMMERMANN, H-J. *Fuzzy Set Theory: And Its Applications.* Kluwer Academic Publishers: London, 2001.