Advances in Fuzzy Systems

Fuzzy Methods and Approximate Reasoning in Geographical Information Systems

Guest Editors: Ferdinando Di Martino, Irina Perfilieva, Salvatore Sessa, and Sabrina Senatore

Fuzzy Methods and Approximate Reasoning in Geographical Information Systems

Advances in Fuzzy Systems

Fuzzy Methods and Approximate Reasoning in Geographical Information Systems

Guest Editors: Ferdinando Di Martino, Irina Perfilieva, Salvatore Sessa, and Sabrina Senatore

Editorial Board

Contents

Hindawi Publishing Corporation
Advances in Fuzzy Systems
Volume 2014, Article ID 840297, 1 page
http://dx.doi.org/10.1155/2014/840297

Editorial

Fuzzy Methods and Approximate Reasoning in Geographical Information Systems

Ferdinando Di Martino,[1] Irina Perfilieva,[2] Salvatore Sessa,[1] and Sabrina Senatore[3]

[1] *Dipartimento di Architettura, Universitá degli Studi di Napoli Federico II, Via Toledo 402, 80134 Napoli, Italy*
[2] *Centre of Excellence IT4Innovations, Institute for Research and Applications of Fuzzy Modelling, University of Ostrava, 30. dubna 22, 70103 Ostrava, Czech Republic*
[3] *Dipartimento di Informatica, Universitá degli Studi di Salerno, Via Ponte don Melillo, Fisciano, 80084 Salerno, Italy*

Correspondence should be addressed to Ferdinando Di Martino; fdimarti@unina.it

Received 2 February 2014; Accepted 2 February 2014; Published 12 March 2014

This issue has been dedicated to the usage of fuzzy logic in the context of geographical information systems (GIS) and received the following papers whose contents are described below.

(i) In the paper of A. Hofmann et al., the authors use a GIS tool which is useful to study the influence of geographic and climatic factors on the terrain passability of armed forces and the integrated rescue system.

(ii) In the paper of F. Di Martino and S. Sessa, the authors propose the usage of the well-known extended Gustafson-Kessel clustering method, encapsulated in a GIS tool, for detecting hotspots in spatial analysis. The data consist of georeferenced patterns corresponding to positions of Taliban's attacks against civilians and soldiers in Afghanistan that happened during the period from 2004 to 2010: the formation through time of new hotspots is observed as well.

(iii) In the paper of M. Burda et al., an application of the so-called fuzzy GUHA method is presented for good peak prediction which was used in order to mine for fuzzy association rules expressed in natural language. The provided data was firstly extended by a creation of artificial variables describing various features of the data. The resulting variables were later on translated into fuzzy GUHA tables with help of evaluative linguistic expressions in order to mine for associations. The found associations were interpreted as fuzzy IF-THEN rules and used jointly with the perception-based logical deduction inference method to predict expected time shift of low rate peaks forecasted by the given physical model.

(iv) In another paper of F. Di Martino and S. Sessa, a fuzzy process for evaluating the reliability of a spatial database is defined: the area of study is partitioned in iso-reliable zones, defined as homogeneous zone in terms of data quality and environmental characteristics. This spatial database includes thematic datasets which in turn includes a set of layers. We estimate the reliability of each thematic dataset and therefore the overall reliability of the spatial database. This method is tested on the spatial dataset of the town of Cava dè Tirreni (Italy) by means of a suitable GIS.

(v) In the paper of F. Di Martino et al., an application of the extended fuzzy C-means algorithm for detecting spatial areas with high concentrations of events, tested to study their temporal evolution, is proposed as well. This algorithm is implemented in a GIS tool. The data consist of georeferenced patterns corresponding to the residence of patients in the district of Naples (Italy) to whom a surgical intervention was carried out to the oto-laryngopharyngeal apparatus during the years from 2008 to 2012.

This special issue presents some noteworthy applications of the spatial analysis realized via GIS. Other applications should be desiderable in the sterminated world of GIS. We are aware that the topics do not easily meet desiderata of fuzzy authors; however we are at the beginning of a theory which is very promising from an applicational point of view, mainly in the spatiotemporal evolution of events which are either difficult to evaluate in the future or "fuzzy" for their same nature.

Ferdinando Di Martino
Irina Perfilieva
Salvatore Sessa
Sabrina Senatore

Hindawi Publishing Corporation
Advances in Fuzzy Systems
Volume 2013, Article ID 131875, 10 pages
http://dx.doi.org/10.1155/2013/131875

Research Article

Mining Linguistic Associations for Emergent Flood Prediction Adjustment

Michal Burda, Pavel Rusnok, and Martin Štěpnička

Institute for Research and Applications of Fuzzy Modeling, National Supercomputing Center IT4Innovations, Division of University of Ostrava, 30. dubna 22, 701 03 Ostrava, Czech Republic

Correspondence should be addressed to Martin Štěpnička; martin.stepnicka@osu.cz

Received 13 October 2013; Accepted 19 October 2013

Academic Editor: Salvatore Sessa

Floods belong to the most hazardous natural disasters and their disaster management heavily relies on precise forecasts. These forecasts are provided by physical models based on differential equations. However, these models do depend on unreliable inputs such as measurements or parameter estimations which causes undesirable inaccuracies. Thus, an appropriate data-mining analysis of the physical model and its precision based on features that determine distinct situations seems to be helpful in adjusting the physical model. An application of fuzzy GUHA method in flood peak prediction is presented. Measured water flow rate data from a system for flood predictions were used in order to mine fuzzy association rules expressed in natural language. The provided data was firstly extended by a generation of artificial variables (features). The resulting variables were later on translated into fuzzy GUHA tables with help of Evaluative Linguistic Expressions in order to mine associations. The found associations were interpreted as fuzzy IF-THEN rules and used jointly with the Perception-based Logical Deduction inference method to predict expected time shift of flow rate peaks forecasted by the given physical model. Results obtained from this adjusted model were statistically evaluated and the improvement in the forecasting accuracy was confirmed.

1. Introduction

Disaster management is generally becoming more and more important task. Among many natural disasters, floods are the one of the most hazardous, and moreover, one of the most frequently occurring in the region of the central Europe. Researchers invest enormous efforts into investigation of distinct flood models that would help to forecast floods and thus provide the disaster management with a reliable decision support that could be used in order to prevent further deceases and material costs.

One of such long-term researches focusing on the disaster management and especially on modeling and forecasting floods gave rise to the creation of the FLOREON, a system for emergent flood prediction [1]. No matter how sophisticated the system is, due to the natural imprecision in data sources (e.g., measuring stations) and due to the natural imprecision in parameter setting (crisp values determined by an expert decision), and having in mind how complicated the whole problem is, it necessarily provides forecasts that are not always precise.

Therefore, it seems to be appropriate to focus on some analysis of the performance of the system that could give at least a vague idea under which conditions the system works, under which conditions it provides us with a certain imprecision, and under which conditions we are able to correct the forecast. Based on the sources of the imprecision, it seems that an appropriate data-mining technique that would involve fuzziness might provide us with promising results and is worth of being attempted. In this investigation, we face the above foreshadowed problem with the help of the fuzzy GUHA method, that is, a specific variant of associations mining technique that allows using the concepts of fuzzy logic in a broad sense [2].

1.1. Brief Problem Description. The data being analyzed come from the measures of water flow rate of the Odra River in Ostrava, Czech Republic. Measuring stations provide us with a flow rate [m^3/s] on hourly basis. The goal is to forecast a future flow rate. This is done by the so-called Math-1D model [3] developed for the FLOREON disaster management IT system [1].

The Math-1D model is a differential equation based model of the flow rate. In order to provide us with flow rate forecasts, it uses information about precipitations (past and forecasted), soil type, river bank shape, and other parameters. Although it is a well-established physical model that is empirically examined, it is not sufficiently reliable. The reason does not lie in the model but in the fact that most of the parameters and input data are highly imprecise. For example, the soil type is provided by a hydrologist expert but, due to natural limitations, without a deeper geological analysis and, moreover, the provided soil type is the same for the whole river flow.

Having in mind these limitations, the Math-1D model forecasts depend mainly on the measured past precipitations and flow rates and on the forecasted precipitations. Thus, provided forecast, though often reliable, may be even highly imprecise. The imprecision may be viewed in two perspectives: in the vertical one and the horizontal one. The vertical imprecision actually means either the overestimation or even worse the underestimation of the flow rate in the culminating peak. For our investigation, the second, that is, the horizontal imprecision, is crucial. That is, we focus on the precision in terms of time; that is, we focus on the question whether and under which conditions the model forecasts the peak discharge earlier or later and how big is the time shift of the peak.

The vertical as well as horizontal imprecision may be significant. As one can see from an exemplary forecast in Figure 1, the real culminating peak can appear a few hours sooner than forecasted. Let us note that the Math-1D model does not use the knowledge of the water flow rate in the past and depends mainly on the precipitations. This explains why it may happen that the model does not fit well the past data (from −119th hour to 0th; see Figure 1). On the other hand, precipitations are rather precise compared to the data from the measuring stations that may not be well calibrated or, even worse, the measuring station may be partially damaged or even fully destroyed (occasionally, it happens that even during a massive flood, measuring stations provide a zero flow rate measurements).

Our task is to analyze and forecast the peak shift on the horizontal (time) axis. In other words, the task is to build a model that would (based on the flow rate measurements and the Model-1D performance in the past) provide disaster management with a valuable information about possible horizontal imprecision of the Math-1D model and, moreover, that would additionally provide the disaster management with an estimation about the peak shift. This peak shift estimation could be used in the corrections of the forecasts.

2. Theoretical Background

In this Section, we introduce fundamental theoretical background that is used in our investigation. As there is no space to introduce all the theoretical concepts in detail, we will provide readers only with a brief introduction and refer to further valuable sources [4–7].

2.1. Evaluative Linguistic Expressions. One of the main constituents of systems of fuzzy/linguistic IF-THEN rules is *evaluative linguistic expressions* [4], in short *evaluative expressions*, for example, *very large*, *more or less hot,* and so forth. They are special expressions of natural language that are used whenever it is important to evaluate a decision situation, to specify the course of development of some process, and in many other situations. Note that their importance and the potential to model their meaning mathematically have been pointed out by Zadeh (e.g., in [8, 9]).

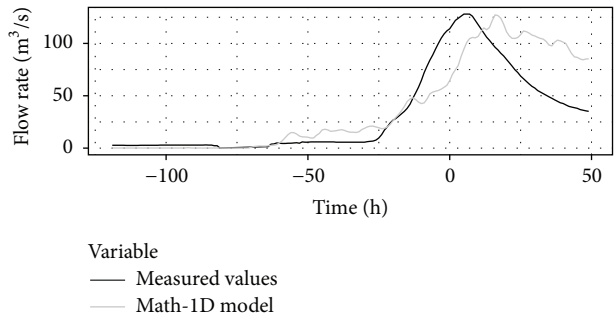

Figure 1: Particular example of real measured values (in black) and the Math-1D model simulation (in gray). Flow rate values [m^3/s] on the vertical axis are measured on a particular measuring station placed on the Odra River starting from time −119 up to 0 (horizontal axis, time [h]). Starting from the time point "0" to the right, the gray curve denotes the forecast of future flow rate values.

Table 1: Linguistic hedges and their abbreviations.

Narrowing effect	Widening effect
Very (Ve)	More or less (ML)
Significantly (Si)	Roughly (Ro)
Extremely (Ex)	Quite roughly (QR)
—	Very roughly (VR)

A simple form of evaluative expressions keeps the following structure:

$$\langle \text{ling} \cdot \text{hedge} \rangle \, \langle \text{atomic evaluative expression} \rangle . \quad (1)$$

Atomic evaluative expressions comprise any of the *canonical* adjectives *small*, *medium*, and *big*, abbreviated in the following as Sm, Me, and Bi, respectively.

Linguistic hedges are specific adverbs that make the meaning of the atomic expression more or less precise. We may distinguish hedges with *narrowing effect*, for example, *very*, *extremely*, and so forth and with *widening effect*, for example, *roughly*, *more or less* and so forth. In the following text, we, without loss of generality, use the hedges introduced in Table 1 that were successfully used in real applications [10] and that are implemented in the LFLC software package [11]. As a special case, the ⟨linguistic hedge⟩ can be empty. Note that our hedges are of so-called inclusive type [12], which means that extensions of more specific evaluative expressions are included in less specific ones; see Figure 2.

Evaluative expressions of the form (1) will generally be denoted by script letters $\mathscr{A}$, $\mathscr{B}$, and so forth. They are used to evaluate values of some variable X. The resulting expressions are called *evaluative linguistic predications* and have the form

$$X \text{ is } \mathscr{A}. \quad (2)$$

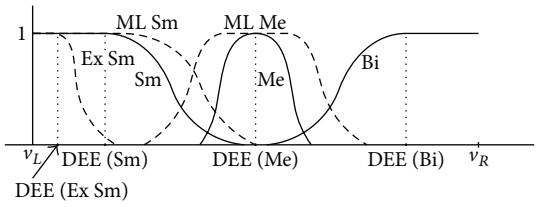

Figure 2: Shapes of extensions (fuzzy sets) of evaluative linguistic expressions. DEE denotes the defuzzified values obtained using the *Defuzzification of Evaluative Expressions*.

Examples of evaluative predications are "temperature is very high," "price is low," and so forth. The model of the meaning of evaluative expressions and predications makes distinction between *intensions* and *extensions* in various *contexts*. The context characterizes a range of possible values. This range can be characterized by a triple of numbers $\langle v_L, v_M, v_R\rangle$, where $v_L, v_M, v_R \in \mathbb{R}$ and $v_L < v_M < v_R$. These numbers characterize the minimal, middle, and maximal values, respectively, of the evaluated characteristics in the specified context of use. Therefore, we will identify the notion of context with the triple $w = \langle v_L, v_M, v_R\rangle$. By $v \in w$ we mean $v \in [v_L, v_R]$. In the sequel, we will work with a set of contexts $W \subset \{\langle v_L, v_M, v_R\rangle \mid v_L, v_M, v_R \in \mathbb{R}, v_L < v_M < v_R\}$ that are given in advance.

The intension of an evaluative predication "X is $\mathscr{A}$" is a certain formula whose interpretation is a function:

$$\text{Int}\,(X \text{ is } \mathscr{A}) : W \longrightarrow \mathscr{F}\,(\mathbb{R})\,; \tag{3}$$

that is, it is a function that assigns a fuzzy set to any context from the set W.

Given an intension (3) and a context $w \in W$, we can define the extension of "X is $\mathscr{A}$" in the context w as a fuzzy set:

$$\text{Int}\,(X \text{ is } \mathscr{A})\,(w) \subsetneq [v_L, v_R]\,, \tag{4}$$

where $\subsetneq$ denotes the relation of fuzzy subsethood.

Convention 1. For the sake of brevity and simplicity and having in mind that an extension is a fuzzy set on a given context, we will omit the notion of extension from our consideration when appropriate and write only the abbreviated forms:

$$\begin{aligned} A &:= (\text{Int}\,(X \text{ is } \mathscr{A})\,(w))\,, \quad w \in W, \\ A\,(v_0) &:= (\text{Int}\,(X \text{ is } \mathscr{A})\,(w))\,(v_0)\,, \quad v_0 \in w, \end{aligned} \tag{5}$$

if there is no danger of any confusion caused by the fact that the left-hand side does not explicitly mention the chosen context w and variable X.

2.2. Linguistic Descriptions. Evaluative predications occur in conditional clauses of natural language of the form

$$\mathscr{R} := \text{IF } X \text{ is } \mathscr{A} \text{ THEN } Y \text{ is } \mathscr{B}, \tag{6}$$

where $\mathscr{A}$, $\mathscr{B}$ are evaluative expressions. The linguistic predication "X is $\mathscr{A}$" is called the *antecedent* and "Y is $\mathscr{B}$" is called the *consequent* of rule (6). Of course, the antecedent may consist of more evaluative predications, joined by the connective "AND." The clauses (6) will be called fuzzy/linguistic IF-THEN rules in the sequel.

Fuzzy/linguistic IF-THEN rules are gathered in a *linguistic description*, which is a set $\text{LD} = \{\mathscr{R}_1, \ldots, \mathscr{R}_m\}$, where

$$\begin{aligned} \mathscr{R}_1 &:= \text{ IF } X \text{ is } \mathscr{A}_1 \text{ THEN } Y \text{ is } \mathscr{B}_1, \\ &\vdots \\ \mathscr{R}_m &:= \text{ IF } X \text{ is } \mathscr{A}_m \text{ THEN } Y \text{ is } \mathscr{B}_m. \end{aligned} \tag{7}$$

Because each rule in (7) is taken as a specific *conditional sentence of natural language*, a linguistic description can be understood as a *specific kind of a (structured) text*. This text can be viewed as a *model* of specific behavior of the system in concern.

The *intension of a fuzzy/linguistic IF-THEN* rule $\mathscr{R}$ in (6) is a function:

$$\text{Int}\,(\mathscr{R}) : W \times W \longrightarrow \mathscr{F}\,(\mathbb{R} \times \mathbb{R})\,. \tag{8}$$

This function assigns to each context $w \in W$ and each context $w' \in W$ a *fuzzy relation* in $w \times w'$. The latter is an *extension* of (8).

We also need to consider a linguistic phenomenon of topic-focus articulation (cf. [13]), which in the case of linguistic descriptions requires us to distinguish the following two sets:

$$\begin{aligned} \text{Topic}_{\text{LD}} &= \left\{\text{Int}\left(X \text{ is } \mathscr{A}_j\right) \mid j = 1, \ldots, m\right\}, \\ \text{Focus}_{\text{LD}} &= \left\{\text{Int}\left(Y \text{ is } \mathscr{B}_j\right) \mid j = 1, \ldots, m\right\}. \end{aligned} \tag{9}$$

The phenomenon of topic-focus articulation plays an important role in the inference method called perception-based logical deduction described below.

Convention 2. Besides the above introduced notions of topic and focus, it is sometimes advantageous to introduce the following notation:

$$\text{Topic}^w_{\text{LD}} = \left\{\text{Int}\left(X \text{ is } \mathscr{A}_j\right)(w) \mid j = 1, \ldots, m\right\}, \tag{10}$$

which will denote the set of extensions of evaluative predications that are contained in Topic_{LD} knowing the particular context w. This notation will be used later on when defining the function of local perception. In the view of Convention 1 one can also easily introduce the $\text{Topic}^w_{\text{LD}}$ as follows:

$$\text{Topic}^w_{\text{LD}} = \left\{A_j \mid j = 1, \ldots, m\right\}. \tag{11}$$

2.3. Ordering of Linguistic Predications. To be able to state relationships among evaluative expressions, for example, when one expression "covers" another, we need an ordering relation defined on the set of them. Let us start with the

ordering on the set of linguistic hedges. We may define the ordering $\leq_{\mathrm{H}}$ of examples of hedges as follows:

$$\mathrm{Ex} \leq_{\mathrm{H}} \mathrm{Si} \leq_{\mathrm{H}} \mathrm{Ve} \leq_{\mathrm{H}} \langle \mathrm{empty} \rangle \leq_{\mathrm{H}} \mathrm{ML} \leq_{\mathrm{H}} \mathrm{Ro} \leq_{\mathrm{H}} \mathrm{QR} \leq_{\mathrm{H}} \mathrm{VR}. \tag{12}$$

We extend the theory of evaluative linguistic expressions by the following *inclusion axiom*. Let $\mathrm{Ker}(A)$ denote the kernel of a fuzzy set A. For any w,

$$\begin{aligned} &\mathrm{Int}\left(X \text{ is } \langle \mathrm{hedge} \rangle_i \mathscr{A}\right)(w) \subseteq \mathrm{Int}\left(X \text{ is } \langle \mathrm{hedge} \rangle_j \mathscr{A}\right)(w) \\ &\mathrm{Ker}\left(\mathrm{Int}\left(X \text{ is } \langle \mathrm{hedge} \rangle_i \mathscr{A}\right)(w)\right) \\ &\quad \subset \mathrm{Ker}\left(\mathrm{Int}\left(X \text{ is } \langle \mathrm{hedge} \rangle_j \mathscr{A}\right)(w)\right) \end{aligned} \tag{13}$$

hold for any atomic expression $\mathscr{A}$ under the assumptions $\langle \mathrm{hedge} \rangle_i \leq_{\mathrm{H}} \langle \mathrm{hedge} \rangle_j$, $i \neq j$.

Based on $\leq_{\mathrm{H}}$ we may define an ordering $\leq_{\mathrm{LE}}$ of evaluative expressions. Let $\mathscr{A}_i$, $\mathscr{A}_j$ be two evaluative expressions such that $\mathscr{A}_i := \langle \mathrm{hedge} \rangle_i \mathscr{A}$ and $\mathscr{A}_j := \langle \mathrm{hedge} \rangle_j \mathscr{A}$. Then we write

$$\mathscr{A}_i \leq_{\mathrm{LE}} \mathscr{A}_j \tag{14}$$

if $\mathscr{A} \in \{\mathrm{Sm}, \mathrm{Me}, \mathrm{Bi}\}$ and $\langle \mathrm{hedge} \rangle_i \leq_{\mathrm{H}} \langle \mathrm{hedge} \rangle_j$.

In other words, evaluative expressions of the same type are ordered according to their specificity which is given by the hedges appearing in the expressions. If we are given two evaluative predications with an atomic expression of a different type, we cannot order them by $\leq_{\mathrm{LE}}$.

Finally, we define the ordering $\leq_{(v_0,w)}$ of evaluative predications wrt. a given observation. Let us be given a context $w \in W$, an observation $v_0 \in w$, and two extensions A_i and A_j from the $\mathrm{Topic}^w_{\mathrm{LD}}$. We write $A_i \leq_{(v_0,w)} A_j$ either if $A_i(v_0) > A_j(v_0)$ or if $A_i(v_0) = A_j(v_0)$ and $\mathscr{A}_i \leq_{\mathrm{LE}} \mathscr{A}_j$.

It should be noted that usually the $\mathrm{Topic}_{\mathrm{LD}}$ contains intensions of evaluative predications which are compound by a conjunction of more than one evaluative predication. In other words, we usually meet the following situation:

$$\begin{aligned} \left(X \text{ is } \mathscr{A}_i\right) &:= \left(X_1 \text{ is } \mathscr{A}_{i_1}\right) \mathrm{AND} \cdots \mathrm{AND} \left(X_K \text{ is } \mathscr{A}_{i_K}\right), \\ \left(X \text{ is } \mathscr{A}_j\right) &:= \left(X_1 \text{ is } \mathscr{A}_{j_1}\right) \mathrm{AND} \cdots \mathrm{AND} \left(X_K \text{ is } \mathscr{A}_{j_K}\right). \end{aligned} \tag{15}$$

In this case, the ordering $\leq_{\mathrm{LE}}$ is preserved with respect to the components:

$$\mathscr{A}_i \leq_{\mathrm{LE}} \mathscr{A}_j \text{ if } \mathscr{A}_{i_k} \leq_{\mathrm{LE}} \mathscr{A}_{j_k} \quad \forall\, k = 1, \ldots, K \tag{16}$$

and the extension of the compound linguistic predication is given as follows:

$$A_i(u_1, \ldots, u_K) = \bigwedge_{k=1}^{K} A_{ik}(u_k). \tag{17}$$

Then, the final ordering $\leq_{(v_0,w)}$ is analogous to the one-dimensional one.

2.4. Perception-Based Logical Deduction. *Perception-based Logical Deduction* (abb. PbLD) is a special inference method aimed at the derivation of results based on fuzzy/linguistic IF-THEN rules. A perception is understood as an evaluative expression assigned to the given input value in the given context. The choice of perception depends on the topic of the specified linguistic description. In other words, perception is always chosen among evaluative expressions which occur in antecedents of IF-THEN rules; see [5, 10, 14].

Based on the ordering $\leq_{(v_0,w)}$ of linguistic predications, a special function of *local perception*

$$\mathrm{LPerc}^{\mathrm{LD}} : w \times W^K \longrightarrow \mathscr{P}\left(\mathrm{Topic}_{\mathrm{LD}}\right) \tag{18}$$

assigns to each value $v_0 = [v_1, \ldots, v_K] \in w$ for $w = [w_1, \ldots, w_K] \in W^K$ a subset of intensions minimal wrt. the ordering $\leq_{(v_0,w)}$

$$\begin{aligned} &\mathrm{LPerc}^{\mathrm{LD}}(v_0, w) \\ &\quad = \left\{A_i \mid A_i(v_0) > 0\,, \forall A_j \in \mathrm{Topic}^w_{\mathrm{LD}} : \left(A_j \leq_{(v_0,w)} A_i\right)\right. \\ &\qquad \left. \Longrightarrow \left(\mathscr{A}_j = \mathscr{A}_i\right)\right\}. \end{aligned} \tag{19}$$

Let LD be a linguistic description (7). Let us consider a context $w \in W$ for the variable X and a context $w' \in W$ for Y. Let an observation $X = v_0$ in the context w be given, where $v_0 \in w$. Then, the following *rule of perception-based logical deduction* (r_{PbLD}) can be introduced:

$$r_{\mathrm{PbLD}} : \frac{\mathrm{LPerc}^{\mathrm{LD}}(v_0, w), \mathrm{LD}}{C}, \tag{20}$$

where C is the conclusion which corresponds to the observation in a way described below. Inputs to this inference rule are linguistic description LD and local perception $\mathrm{LPerc}^{\mathrm{LD}}(v_0, w)$ from (19). This local perception is formed by a set of evaluative expressions from antecedents of IF-THEN rules (i.e., from the topic) of the given linguistic description. Formula (19) chooses these antecedents which best fit the given numerical input v_0; in other words, they are the most specific according to the ordering $\leq_{(v_0,w)}$.

Once one or more antecedents $A_{i_\ell} \in \mathrm{Topic}^w_{\mathrm{LD}}$, $i_\ell = 1, \ldots, L$ are chosen according to (19), we compute for any of them conclusions C_{i_ℓ}:

$$C_{i_\ell}(v) = A_{i_\ell}(v_0) \longrightarrow B_{i_\ell}(v), \quad v \in w', \tag{21}$$

where $\rightarrow$ is the Łukasiewicz implication [2] given by $a \rightarrow b = 1 \wedge (1 - a + b)$.

Suppose that $\mathrm{LPerc}^{\mathrm{LD}}(v_0, w)$ is nonempty; that is, $L > 0$. Then the final conclusion C is given as the Gödel intersection of the set of all L conclusions C_{i_ℓ} that correspond to L members in $\mathrm{LPerc}^{\mathrm{LD}}(v_0, w)$; that is,

$$C(v) = \bigwedge_{\ell=1}^{L} C_{i_\ell}(v) = \bigwedge_{\ell=1}^{L} \left(A_{i_\ell}(v_0) \longrightarrow B_{i_\ell}(v)\right). \tag{22}$$

In many application, the inferred output fuzzy set C needs to be defuzzified to a crisp value in w'. For this task, a special defuzzification technique called *Defuzzification of Evaluative Expressions* (abb. DEE) has been proposed. In principle, this defuzzification is a combination of *First-Of-Maxima* (FOM), *Mean-Of-Maxima* (MOM), and *Last-Of-Maxima* (LOM) that are applied based on the classification of the output fuzzy sets. Particularly, if the inferred fuzzy set is of the type `Small` (nonincreasing), the LOM is applied; if the inferred output is of the type `Big` (nondecreasing), the FOM is applied; and finally, if the inferred output is of the type `Medium`, the MOM is applied; see Figure 2.

3. Fuzzy GUHA: Linguistic Associations Mining

In this paper, we employ the so-called linguistic associations mining [15] for the fuzzy rule base identification. This approach, mostly known as mining association rules [16], was firstly introduced as GUHA method [17, 18]. It finds distinct statistically approved associations between attributes of given objects. Particularly, the GUHA method deals with Table 2 where $o_1, \ldots, o_n$ denote objects, $X_1, \ldots, X_q$ denote independent boolean attributes, Z denotes the dependent (explained) boolean attribute, and finally, symbols a_{ij} (or a_i) $\in \{0, 1\}$ denote whether an object o_i carries an attribute X_j (or Z) or not.

The original GUHA allowed only boolean attributes to be involved; see [19]. Since most of the features of objects are measured on the real interval, standard approach assumed categorization of quantitative variables and subsequently definition of boolean variables for every category.

The goal of the GUHA method is to search for linguistic associations of the form

$$C\left(X_1, \ldots, X_p\right) \simeq D\left(Z\right), \tag{23}$$

where C, D are (compound) evaluative predications [20] containing only the connective AND and $X_1, \ldots, X_p$ for $p \leq q$ are all variables occurring in C. The C, D are called the *antecedent* and *consequent*, respectively. Generally, for the GUHA method, the well-known fourfold table is constructed; see Table 3.

Symbol a, in Table 3, denotes the number of positive occurrences of C as well as D; b is the number of positive occurrences of C and of negated D, that is, of "not D." Analogous meaning has the numbers c and d. For our purposes, only numbers a and b are important.

The relationship between the antecedent and the consequent is described by so-called *quantifier* $\simeq$. There are many quantifiers that characterize validity of association (23) in the data [18]. For our task, we use the so-called binary multitudinal quantifier $\simeq := \sqsubset_r^\gamma$. This quantifier is taken as true if

$$\begin{aligned} \frac{a}{a+b} &> \gamma, \\ \frac{a}{n} &> r, \end{aligned} \tag{24}$$

Table 2: Standard GUHA table.

	X_1	$\ldots$	X_q	Z
o_1	a_{11}	$\ldots$	a_{1q}	a_1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
o_n	a_{n1}	$\ldots$	a_{nq}	a_n

Table 3: Classical GUHA fourfold table.

	D	Not D
C	a	b
Not C	c	d

where $\gamma \in [0, 1]$ is a confidence degree and $r \in [0, 1]$ is a support degree.

Example 1. For example, let us consider Table 4.

Depending on the chosen confidence and support degree, the GUHA method could generate, for example, the following linguistic association:

$$C\left(\mathrm{BMI}_{>25}, \mathrm{Ch}_{>6.2}\right) \simeq \mathrm{D}\left(\mathrm{BP}_{>130/90}\right). \tag{25}$$

According to [21], there are two approaches in treating quantitative variables in association rules mining. The first one is to categorize the variables using the predefined concept hierarchies (e.g., $\mathrm{BMI}_{\leq 25}$). And the second one is to search for clusters in a variable and discretize it according to the found clusters (distribution of the data). Nevertheless both approaches divide numerical variables into crisp intervals.

In many situations, including our situation, it is better to define fuzzy sets on the numerical variables and use the fuzzy variant of the GUHA method [15, 22]. In this case we have also two possibilities how to treat quantitative variables. Either we will apply fuzzy clustering or we will use some predefined concepts. Because of the well-developed theory of Evaluative Linguistic Expressions (Section 2.1) we chose the latter approach.

In the fuzzy variant of the method, the attributes are not boolean but rather vague. The minimum (resp., maximum) of a particular attribute becomes v_L (resp., v_R) and thus we obtain the context $\langle v_L, v_M, v_R \rangle$ for the given attribute (v_M might be median, mean, or other value between v_L and v_R). With canonical adjectives Sm, Me, and Bi and seven different linguistic hedges we may define more than 20 fuzzy sets for every quantitative variable (attribute). The values a_{ij} (or a_i) are now elements of the interval $[0, 1]$ that express membership degrees.

For example, instead of defining a boolean variable $\mathrm{BMI}_{\leq 25}$ (see Table 4), we take the quantitative variable BMI and generate all the possible evaluative linguistic predications and define fuzzy sets $\mathrm{BMI}_{\mathrm{ExSm}}, \mathrm{BMI}_{\mathrm{SiSm}}, \mathrm{BMI}_{\mathrm{VeSm}}, \mathrm{BMI}_{\mathrm{Sm}}, \ldots, \mathrm{BMI}_{\mathrm{VRSm}}$, so that the first column in Table 4 is replaced with Table 5. In this way we are able to separate a group of malnourished people ($\mathrm{BMI}_{\mathrm{ExSm}}$). Analogically to capture such cases of people, who have almost ideal BMI index, we define $\mathrm{BMI}_{\mathrm{Me}}$, $\mathrm{BMI}_{\mathrm{RoMe}}$. Finally, instead of $\mathrm{BMI}_{>25}$ we define $\mathrm{BMI}_{\mathrm{ExBi}}$, $\mathrm{BMI}_{\mathrm{SiBi}}$, and so forth. Thus, we

Table 4: Example of GUHA table. $\text{BMI}_{\leq 25}$ denotes Body-Mass-Index lower or equal to 25, $\text{BMI}_{>25}$ denotes the same index above 25, $\text{Chol}_{\leq 6.2}$ denotes cholesterol lower or equal to 6.2, $\text{Chol}_{>6.2}$ denotes cholesterol higher than 6.2, $\text{BP}_{\leq 130/90}$ denotes blood pressure lower or equal to 130/90, and $\text{BP}_{>130/90}$ denotes blood pressure higher than 130/90. Objects o_i are particular patients.

	$\text{BMI}_{\leq 25}$	$\text{BMI}_{>25}$	$\text{Chol}_{\leq 6.2}$	$\text{Chol}_{>6.2}$	$\text{BP}_{\leq 130/90}$	$\text{BP}_{>130/90}$
o_1	1	0	1	0	1	0
o_2	0	1	0	1	0	1
o_3	0	1	1	0	0	1
o_4	1	0	1	0	1	0
o_5	0	1	0	1	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
o_n	0	0	0	1	0	1

Table 5: Example of a part of a fuzzy GUHA table with variable BMI and one canonical adjective Sm. BMI_{ExSm} denotes the fuzzy set "Extremely Small Body-Mass-Index", BMI_{SiSm} denotes the fuzzy set "Significantly Small Body-Mass-Index," and so forth. Objects o_i are particular patients and numbers in the table are their membership degrees in particular fuzzy set.

	BMI_{ExSm}	BMI_{SiSm}	BMI_{VeSm}	BMI_{Sm}	…	BMI_{QRSm}	BMI_{VRSm}
o_1	0.5	0.6	0.7	0.8	…	1	1
o_2	0.8	0.9	1	1	…	1	1
o_3	0	0	0	0.1	…	0.3	0.4
o_4	0	0	0	0	…	0	0
o_5	0.6	0.9	1	1	…	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
o_n	0	0.5	0.8	0.9	…	1	1

add information that was lost by the transition from the quantitative variable BMI into two boolean variables $\text{BMI}_{\leq 25}$ and $\text{BMI}_{>25}$. More importantly, we also capture gradual transitions between different groups of people. An object o_i (in our example a patient) might have membership degree to the fuzzy set BMI_{Sm} equal to $0.4(\text{BMI}_{\text{Sm}}(o_i) = 0.4)$ and simultaneously o_i belongs to the fuzzy set BMI_{Me} with the degree $0.3(\text{BMI}_{\text{Me}}(o_i) = 0.3)$. This way we capture the information about the patient that is on the transition from being underweight to having ideal BMI index. This kind of information cannot be captured by crisp intervals.

In this way we treat every quantitative variable so that the final fuzzy GUHA table will look similarly to Table 6.

The fourfold table analogous to Table 3 is constructed also for the fuzzy variant of the method. The difference is that the numbers a, b, c, and d are not summations of 1 s and 0 s but summations of membership degrees of data into fuzzy sets representing the antecedent C and consequent D or their complements, respectively. Naturally, the fact that antecedent C as well as consequent D holds simultaneously leads to the natural use of a t-norm [23]. In our case, we use the Gödel t-norm that is the minimum operation. For example, if an object o_i belongs to a given antecedent in a degree 0.7 and to a given consequent in a degree 0.6, the value that enters the summation equals to $\min\{0.7, 0.6\} = 0.6$. Summation of such values over all the objects equals the value a in Table 3; the other values from the table are determined analogously. The rest of the ideas of the method remain the same.

By using fuzzy sets, we generally get more precise results, and, more importantly, we avoid undesirable threshold effects [24]. The further advantage is that the method searches for implicative associations that may be directly interpreted as fuzzy rules for the PbLD inference system.

Example 2. A confirmed association as

$$C\left(\text{BMI}_{\text{ExBi}}, \text{Chol}_{\text{VeBi}}\right) \sqsubset_r^{\gamma} D\left(\text{BP}_{\text{MLBi}}\right) \tag{26}$$

may be directly interpreted as the following fuzzy rule:

"*IF Body-Mass-Index is Extremely Big AND Cholesterol level is Very Big THEN Blood Pressure is More or Less Big.*"

This approach has been found very efficient and reasonable, for example, for the identification of the so-called Fuzzy Rule Base Ensemble [25] which is a special ensemble technique for time series forecasting [26] that uses fuzzy rules to determine weights of individual forecasting methods. Naturally, the overlapping of extensions of linguistic expressions causes a massive generation of redundant associations. However, there exist efficient methods that detect and remove these redundancies automatically; see [6, 7].

In Section 4.3, we apply this method to artificial variables computed from the measures of water flow in order to obtain interesting descriptions of water flow rate peak time shift.

4. Data Analysis

4.1. Data Description. As mentioned in the introduction, we are provided only with the data from the measuring stations and from the Math-1D model implemented in the FLOREON system. Unlike the Math-1D model, we are neither provided

Table 6: Example of fuzzy GUHA table (compared with Table 4).

	BMI_{ExSm}	…	$Chol_{ExSm}$	…	$Chol_{ExBi}$	BP_{ExSm}	…	BP_{ExBi}
o_1	0.5	…	0.7	…	0	1	…	0
o_2	0.8	…	1	…	0	0.4	…	0
o_3	0	…	0	…	0.1	0	…	0.4
o_4	0	…	0	…	0.4	0	…	0.3
o_5	0.6	…	1	…	0	1	…	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
o_n	0	…	0.8	…	0	0.5	…	0

with the measured precipitations nor with their forecasts nor with other physical attributes or their estimations. The reason is that this is the domain for the physical model Math-1D and our task is not to build another competitive physical model but to concentrate on the analysis of the existing one. However, in order to deal with the (fuzzy) GUHA method, we need to generate several features (artificial variables) and investigate the question, which of those variables have some influence on the performance of the model.

For the purpose of this investigation, we were provided with the data set collected from different *events* (floods) on the measuring station *Svinov* placed on the Odra River (Svinov is a part of the Ostrava city through which the Odra River flows. Naturally, the measuring station carries the same name). The whole data set is divided into 57 *simulations*. Each simulation captures a state of the system (provided real values and model values) at some time point t that is for each simulation denoted by zero ($t = 0$). Each simulation can be further divided into past and future data measured or simulated on the hourly basis.

So, we can introduce the following two sets:

$$\begin{aligned} \text{Past} &= \{t \mid -119, -118, \ldots, 0\}, \\ \text{Future} &= \{t \mid 1, 2, \ldots, 48\}, \end{aligned} \tag{27}$$

and the two time dependent variables, namely, the real water flow rate at time t and the originally modelled flow rate at time t, denoted by r_t and m_t, respectively. Thus, we can also introduce the following sets:

$$r_{\text{Past}} = \{r_t \mid t \in \text{Past}\}, \qquad r_{\text{Future}} = \{r_t \mid t \in \text{Future}\}, \tag{28}$$

and analogously

$$m_{\text{Past}} = \{m_t \mid t \in \text{Past}\}, \qquad m_{\text{Future}} = \{m_t \mid t \in \text{Future}\}. \tag{29}$$

Indeed, the values r_{Future} had been unknown at the time point $t = 0$ and they were added to the data later on only for the comparison and efficiency evaluation purposes. The values m_{Future} are forecasts made by the original Math-1D model that were at disposal at the time point $t = 0$.

The aim is to analyze associations between input variables that were at disposal at the time $t = 0$ (r_{Past}, m_{Past}, and m_{Future}) and the dependent variable which was (for this stage of investigation) chosen to be the *peak-time* r_{Peak}, that is, the time of maximum water flow rate:

$$r_{\text{Peak}} = \arg \max_{t \in \text{Future}} r_t. \tag{30}$$

For the sake of result quality evaluation, the data was split into a training set and a testing set in the ratio of 2 : 1, that is, 38 simulations for the training and 19 simulations for the testing.

4.2. Features Generation and Reduction. For each simulation s, a set of features was extracted by applying several statistical characteristics on different vectors of data that were derived from r_{Past} and m_{Past}. Namely, the following statistics were utilized: mean $\overline{u}$, standard deviation σ_u, median $\widetilde{u}$, minimum $\min(u)$, maximum $\max(u)$, range $(\max(u) - \min(u))$, interquartile range $(q_{0.75}(u) - q_{0.25}(u))$, difference of the last value and the mean $(u_0 - \overline{u})$, coefficient of variation CV_u, difference of the mean and the median $(\overline{u} - \widetilde{u})$, absolute difference of the mean and the median $|\overline{u} - \widetilde{u}|$, skewness $\text{Skew}(u)$ and its absolute value $|\text{Skew}(u)|$, kurtosis $\text{Kurt}(u)$ and its absolute value $|\text{Kurt}(u)|$, slope β_u computed from linear regression of $u_t = \iota_u + \beta_u t + \varepsilon_u$ (where ι_u is the intercept and ε_u is the residual error), and trend strength $\text{Trend}(u)$ computed as a P value of the hypothesis $\beta_u = 0$.

All the statistics listed above were computed for each of the following data $u \in \{r_{\text{Past}}, m_{\text{Past}}\}$. Additionally, the same statistics were determined for the following further newly created data vectors:

$$\begin{aligned} \log(u) &= \{\log|u_t| \mid t \in \{-119, -118, \ldots, 0\}\}, \\ \text{diff} &= \{u_{t+1} - u_t \mid t \in \{-119, -118, \ldots, -1\}\}, \\ \log(\text{diff}) &= \{\log|u_{t+1} - u_t| \mid t \in \{-119, -118, \ldots, -1\}\}, \end{aligned} \tag{31}$$

where again $u \in \{r_{\text{Past}}, m_{\text{Past}}\}$.

Analogously, the same statistics have been utilized also for $u = m_{\text{Future}}$ with the only difference stemming from the different time values; that is, they were applied to

$$\begin{aligned} \log(m) &= \{\log|m_t| \mid t \in \{1, 2, \ldots, 48\}\}, \\ \text{diff}_m &= \{m_{t+1} - m_t \mid t \in \{2, 3, \ldots, 48\}\}, \\ \log(\text{diff}_m) &= \{\log|m_{t+1} - m_t| \mid t \in \{2, 3, \ldots, 48\}\}. \end{aligned} \tag{32}$$

Finally, the time point of the forecasted peak,

$$m_{\text{Peak}} = \arg \max_{t \in \text{Future}} m_t, \tag{33}$$

was also added as an additional feature. It means that a total amount of 205 new features were generated.

Table 7: Example of a part of the fuzzy GUHA table for peak shift of PS of the peak forecasted by the Math-1D model. Objects s_i are particular simulations.

	$\sigma_{\log(r_{\text{Past}})\,\text{ExSm}}$	...	$m_{\text{Peak ExBi}}$	PS_{ExSm}	...	PS_{ExBi}
s_1	0.97	...	0.62	0.45	...	0
s_2	0	...	0.2	0	...	0.58
s_3	0.75	...	0.97	0.66	...	0
⋮	⋮	⋮	⋮	⋮	⋮	
s_{38}	0.66	...	0.74	0.69	...	0

Table 8: Examples of fuzzy rules found by the fuzzy GUHA method.

Rule	IF part			THEN part
	$\sigma_{\log(r_{\text{Past}})}$	$\text{CV}_{\log(r_{\text{Past}})}$	m_{Peak}	PS
$\mathscr{R}_1$	Sm	VR Sm	Ve Bi	QR Sm
$\mathscr{R}_2$	VR Sm	Ve Sm	Ve Bi	Ro Sm
$\mathscr{R}_3$	Ve Sm	—	Ex Bi	ML Sm
...	...	...	...	...
$\mathscr{R}_{69}$	VR Me	QR Sm	VR Bi	Ro Me

From the pool of features, a regression method [27] was utilized to select those, which had the highest significance for a regression model. Particularly, the dependent variable

$$\text{PS} = \left(r_{\text{Peak}} - m_{\text{Peak}}\right), \tag{34}$$

denoting the peak shift, was modelled with the linear regression of all the generated features. After that, statistical significance of all the regression coefficients was tested and only features with P-value below 0.05 were selected.

In this way, we ended the feature selection with the following three features: $\sigma_{\log(r_{\text{Past}})}$: standard deviation of $\log(r_{\text{Past}})$; $CV_{\log(r_{\text{Past}})}$: coefficient of variation of $\log(r_{\text{Past}})$; and finally, m_{Peak}: time point of the forecasted peak given by (33).

4.3. Fuzzy GUHA Application. All computed features, which were found statistically significant, as described in the previous subsection, are viewed as quantitative variables. In order to use them in mining linguistic associations, we had to convert them into fuzzy attributes. More specifically, we generated all the possible linguistic expressions (see Section 2) and determined appropriate contexts for each of the variables, and finally, for each simulation, we determined the degrees of membership of the given simulation into the extensions of the linguistic expressions for each variable. Such process turned the three antecedent variables into 63 fuzzy attributes—each related certain evaluative linguistic expression (21 expressions for each variable; see Section 2.1).

The above introduced variable $\text{PS} = (r_{\text{Peak}} - m_{\text{Peak}})$ is the dependent variable whose dependence on the generated attributes appearing in antecedents is being "explained" with the help of the fuzzy GUHA method and the generated linguistic associations; see Section 3.

Part of the resulting fuzzy GUHA table that contained 84 columns, 63 for antecedent attributes, and 21 for consequent attributes is shown in Table 7.

Upon the choice of the multitudinal implicational quantifier and the degree of confidence $\gamma = 0.7$ and the degree of support $r = 0.1$, the fuzzy GUHA generated considerably many linguistic associations. After the application of the redundancy detection and removal algorithm [7], we have obtained 69 fuzzy rules Table 8 that have a twofold importance:

(i) they describe the situations, under which the disaster management may expect some time shift of the water flow rate peak, which is essential for precise warning and evacuation of people or other preparations works that may save material costs of the approaching disaster;

(ii) connected to the PbLD inference mechanism, they may be directly used to forecast the time shift of the peak originally forecasted by the Math-1D model and, thus, to directly correct and precisiate the forecast by the physical model.

5. Prediction, Results, and Evaluation

5.1. Results and Evaluation. The prediction model was evaluated on a testing dataset, that is, on data previously hidden during the whole data-mining procedure. The testing dataset consists of 19 simulations, each simulation containing hourly flow rates for five days in the past and two days of predictions for the future.

On the testing simulations, the prediction accuracy of the time of culminating-peak was compared between the original Math-1D model and the Math-1D model newly adjusted with GUHA association rules.

For each testing simulation s and model—either original (by Math-1D) or adjusted (with the help of fuzzy rules)—a prediction error $e(s)$ was evaluated as follows:

$$e\left(s\right) = m_{\text{Peak}}\left(s\right) - r_{\text{Peak}}\left(s\right), \tag{35}$$

Table 9: A comparison of peak forecast errors $e_{m,s}$ among the original and the adjusted model. The table presents basic statistics computed from $e_{m,s}$ evaluated on testing data.

Model	Min	1st quart.	Mean	Stdev	Median	3rd quart.	Max
Original	−0.33	0.31	0.603	0.52	0.46	0.88	1.92
Adjusted	−1.33	−0.69	−0.205	0.65	−0.12	0.07	1.02

Table 10: Statistical tests of hypotheses of $\bar{e}_m = 0$: Both Wilcoxon signed rank sum test and one sample t-test indicate the error e_{original} to be very likely nonzero. For adjusted model, the variability of e_{adjusted} can be explained by randomness.

Model	Wilcoxon test		One sample t-test		
	V	P value	t	df	P value
Original	166	0.0004871	5.042	18	0.00008486
Adjusted	61	0.1776	−1.373	18	0.1866

where $m_{\text{Peak}}(s)$ is the peak time forecasted for simulation s by a given model and $r_{\text{Peak}}(s)$ is the time of real occurrence of the peak in simulation s; see also formulas (30)–(33). Summary of the comparison can be found in Table 9.

Briefly, it can be stated that the original model expects the flood peaks approximately a half an hour later than in reality, on the testing dataset. After adjustments made by our GUHA model, the estimates become more accurate. More precisely, the original (Math-1D) model error is on average 0.603 days (with standard deviation 0.521). The error of the adjusted model is −0.205 days (with standard deviation 0.65).

A bias towards positive values of the original model was also justified by the one sample Wilcoxon rank sum test [28]: a null hypothesis of zero shift was rejected with P value = 0.000487, on the original model. On the other hand, the same hypothesis cannot be rejected for the adjusted model (with P value = 0.1776). Similar results were also obtained with the one sample t-test (see Table 10).

6. Conclusion

In this paper, we attempted to deal with an adjustment of a physical model of water flow rate during floods with the help of linguistic associations mining. As any physical model based on differential equations (the Math-1D model, in our case) is highly dependent on many unreliable parameters, it seems reasonable to perform some real data analysis that would inform us, when and under which conditions the model is (in terms of the culminating water flow rate peak) time lagged or vice-versa too much ahead.

We approached the task with the help of the fuzzy GUHA method that automatically generates linguistic associations. The provided data was firstly extended by a creation of artificial variables describing various features of the data. The resulting variables were later on translated into fuzzy GUHA table using the so-called Evaluative Linguistic Expressions. This table was used to mine the associations that may be directly interpreted as fuzzy IF-THEN rules. Such interpretation is beneficial not only because of its interpretability but it can also be used jointly with the Perception-based Logical Deduction inference method in order to predict expected time shift of the flood peaks originally forecasted by the physical model. Results obtained from this adjusted model were statistically evaluated in order to confirm the improvement in the forecasting accuracy.

Let us note that the data-mining analysis as well as experimental evaluation was performed only on a single measuring station Svinov placed on Odra River. Indeed, as the physical model depends on many imprecise and estimated parameters that may differ over the river flow, each station would require its own analysis. However, as the number of stations in the whole region is rather low (9 stations placed on four main rivers), such approach is obviously feasible. Thus the promising results give chance for further and deeper analysis that could enhance the disaster management by more accurate physical models with forecasts adjusted by fuzzy IF-THEN rules. On the other hand, there is a serious complication in the lack of the past data that could be analyzed. The high number of previous floods is unfortunately not accompanied by a sufficiently high number of precise data. As we have mentioned, there was, for example, a problem of measured zero water flow rates even during massive floods due to uncalibrated measuring stations or due to other unspecified reasons. This lack of reliable data may significantly complicate the situation.

As the first step for future research, we plan to extend our investigation by using measured past precipitations and possibly also the forecasted future precipitations that are already at disposal to the Math-1D model but that were not at disposal to our data analysis presented in this paper.

Acknowledgment

This work was supported by the European Regional Development Fund in the IT4Innovations Centre of Excellence Project (CZ.1.05/1.1.00/02.0070).

References

[1] J. Martinovič, S. Štolfa, J. Kožusznik, J. Unucka, and I. Vondrák, "FLOREON—the system for an emergent flood prediction," in *ECECFUBUTEC- EUROMEDIA*, Porto, Portugal, 2008.

[2] V. Novák, I. Perfilieva, and J. Močkoř, *Mathematical Principles of Fuzzy Logic*, Kluwer Academic Publishers, Boston, Mass, USA, 1999.

[3] P. Kubíček and T. Kozubek, "Mathematical-analytical solutions of the flood wave and its use in practice," (in czech). VŠB-TU Ostrava, 150.

[4] V. Novák, "A comprehensive theory of trichotomous evaluative linguistic expressions," *Fuzzy Sets and Systems*, vol. 159, no. 22, pp. 2939–2969, 2008.

[5] V. Novák and A. Dvořák, "Formalization of commonsense reasoning in fuzzy logic in broader sense," *Applied and Computational Mathematics*, vol. 10, no. 1, pp. 106–121, 2011.

[6] A. Dvořák, M. Štěpnička, and L. Vavříčková, "Redundancies in systems of fuzzy/linguistic if-then rules, pages," in *Proceedings of the 7th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT '11)*, Advances in Intelligent Systems Research, pp. 1022–1029, Atlantic Press, Paris, France, 2011.

[7] L. Štěpničková, M. Štěpnička, and A. Dvořák, "New results on redundancies of fuzzy/linguistic if-then rules," in *Proceedings of the 8th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT '13)*, pp. 400–407, Atlantic Press, Milano, Italy.

[8] L. A. Zadeh, "Precisiated natural language (PNL)," *AI Magazine*, vol. 25, no. 3, pp. 74–91, 2004.

[9] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning—I," *Information Sciences*, vol. 8, no. 3, pp. 199–249, 1975.

[10] M. Štěpnička, A. Dvořák, V. Pavliska, and L. Vavříčková, "A linguistic approach to time series modeling with the help of F-transform," *Fuzzy Sets and Systems*, vol. 180, no. 1, pp. 164–184, 2011.

[11] A. Dvořák, H. Habiballa, V. Novák, and V. Pavliska, "The concept of LFLC 2000—its specificity, realization and power of applications," *Computers in Industry*, vol. 51, no. 3, pp. 269–280, 2003.

[12] M. De Cock and E. E. Kerre, "Fuzzy modifiers based on fuzzy relations," *Information Sciences*, vol. 160, no. 1–4, pp. 173–199, 2004.

[13] E. Hajičová, B. Partee, and P. Sgall, *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*, Kluwer, Dordrecht, The Netherlands, 1998.

[14] V. Novák, "Perception-based logical deduction," in *Computational Intelligence, Theory and Applications, Advances in Soft Computing*, B. Reusch, Ed., pp. 237–250, Springer, Berlin, Germany, 2005.

[15] J. Kupka and I. Tomanová, "Some extensions of mining of linguistic associations," *Neural Network World*, vol. 20, no. 1, pp. 27–44, 2010.

[16] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proceedings of the 20th International Conference on Very Large Databases*, pp. 487–499, AAAI Press, Santiago de Chile, Chile, 1994.

[17] P. Hájek, "The question of a general concept of the GUHA method," *Kybernetika*, vol. 4, pp. 505–515, 1968.

[18] P. Hájek and T. Havránek, *Mechanizing Hypothesis Formation: Mathematical Foundations for a General Theory*, Springer, Berlin, Germany, 1978.

[19] P. Hájek, M. Holeňa, and J. Rauch, "The GUHA method and its meaning for data mining," *Journal of Computer and System Sciences*, vol. 76, no. 1, pp. 34–48, 2010.

[20] V. Novák, "A comprehensive theory of trichotomous evaluative linguistic expressions," *Fuzzy Sets and Systems*, vol. 159, no. 22, pp. 2939–2969, 2008.

[21] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2006.

[22] V. Novák, I. Perfilieva, A. Dvořák, G. Chen, Q. Wei, and P. Yan, "Mining pure linguistic associations from numerical data," *International Journal of Approximate Reasoning*, vol. 48, no. 1, pp. 4–22, 2008.

[23] E. P. Klement, R. Mesiar, and E. Pap, *Triangular Norms*, vol. 8 of *Trends in Logic*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.

[24] T. Sudkamp, "Examples, counterexamples, and measuring fuzzy associations," *Fuzzy Sets and Systems*, vol. 149, no. 1, pp. 57–71, 2005.

[25] L. Štěpničková, M. Štěpnička, and D. Sikora, "Fuzzy rule-based ensemble with use linguistic associations mining for time series prediction," in *Proceedings of the 8th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT '13)*, pp. 408–415, Atlantic Press, Milano, Italy, 2013.

[26] D. Sikora, M. Štěpnička, and L. Vavříčková, "Fuzzy rule-based ensemble forecasting: Introductory study," in *Synergies of Soft Computing and Statistics for Intelligent Data Analysis*, vol. 190 of *Advances in Intelligent Systems and Computing*, pp. 379–387, Springer, 2013.

[27] J. M. Chambers, A. Freeny, and R. M. Heiberger, "Statistical models in S," in *Linear Models*, Wadsworth & Brooks/Cole, 1992.

[28] M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods*, A Wiley Publication in Applied Statistics, Wiley, New York, NY, USA, 1973.

Hindawi Publishing Corporation
Advances in Fuzzy Systems
Volume 2013, Article ID 107358, 9 pages
http://dx.doi.org/10.1155/2013/107358

Research Article

Fuzzy Reliability in Spatial Databases

Ferdinando Di Martino[1] and Salvatore Sessa[1,2]

[1] *Università degli Studi di Napoli Federico II, Dipartimento di Architettura, Via Monteoliveto 3, 80134 Napoli, Italy*
[2] *Università degli Studi di Napoli Federico II, Centro Interdipartimentale per l'Analisi e la Progettazione Urbana Luigi Pisciotti, Via Toledo 402, 80134 Napoli, Italy*

Correspondence should be addressed to Salvatore Sessa; sessa@unina.it

Received 10 October 2013; Accepted 27 October 2013

Academic Editor: Sabrina Senatore

Today it is very difficult to evaluate the quality of spatial databases, mainly for the heterogeneity of input data. We define a fuzzy process for evaluating the reliability of a spatial database: the area of study is partitioned in isoreliable zones, defined as homogeneous zones in terms of data quality and environmental characteristics. We model a spatial database in thematic datasets; each thematic dataset concerns a specific spatial domain and includes a set of layers. We estimate the reliability of each thematic dataset and therefore the overall reliability of the spatial database. We have tested this method on the spatial dataset of the town of Cava de' Tirreni (Italy).

1. Introduction

Fuzzy rule-based models are applied in geographical information systems (GIS) [1–3] and we use our previous approach [4, 5] for estimating the reliability of spatial databases. There the concept of geodata "reliability" was introduced as a fuzzy measure of the quality of geodata, based on the analysis of uncertainty and quality of the data. Strictly speaking, in [5] the authors implement a tool called (Fuzzy Spatial Reliability Analysis) Fuzzy-SRA [6] for studying the reliability of the intrinsic vulnerability of aquifers by utilizing the DRASTIC model, encapsulated in a GIS; in [4] Fuzzy-SRA is used for estimating the reliability of the aerophotogrammetric set of geographic layers of the island of Procida (near Naples, Italy) and in [7] Fuzzy-SRA is applied in a GIS tool for implementing a fuzzy rule-based system for analyzing the eruption risk of the famous vulcan Vesuvius.

As the first step, we need to divide the geographic area of study in isoreliable zones, that is, in zones having (quasi) homogeneous data quality and geographical characteristics. An expert sets the characteristics related to the quality of each layer (e.g., the percent of uncoded spot elevation features). Each characteristic, called "parameter," is a measurable entity that could affect the quality of the dataset. After calculating the value of a parameter, a fuzzification process is applied for estimating the quality of the set of layers, where each fuzzy set is given by a triangular fuzzy number (TFN), which in turn is identified from a linguistic label. In other words, an isoreliable zone is a subarea of the area of study in which the quality of the geodata is homogeneous; that is, the values of the parameters are similar and with the same geographical characteristics (e.g., a flat country).

The expert creates a fuzzy partition, labelling the TFNs with linguistic labels (say) $\alpha_1, \alpha_2, \ldots, \alpha_n$ (see, e.g., Table 3) for each parameter. An isoreliable zone is associated with the linguistic label of the corresponding TFN, for which the membership degree of the parameter is the highest one. This process is iterated for each parameter of each layer. The elements of Fuzzy-SRA, associated with each parameter, are fuzzy attributes represented by a string. To clarify how these strings are composed, now we suppose, as an example, that we have partitioned the area of study in 5 isoreliable zones, $O1, \ldots, O5$; hence we create a fuzzy partition of the domain discourse of the parameter A in 6 fuzzy sets with linguistic labels, respectively, $\alpha_1, \alpha_2, \ldots, \alpha_6$. After the

Table 1: Example of association of isoreliable zones with linguistic labels.

Isoreliable zone	Linguistic label
$O1$	α_2
$O2$	α_1
$O3$	α_2
$O4$	α_6
$O5$	α_4

Table 2: Spatial datasets and related attributes used in [4].

Layer	Parameter
Spot elevations (SE)	Density of uncoded SE
	Density of coded SE per ha
Contour lines	Mean density of contour lines for ha/mean slope
	Standard deviation of contour lines for ha/mean slope
Buildings	Mean of the absolute value of perimeter-shape length
	Mean of the absolute value of shape area
Network streets	Mean density of buffer area covered by buildings
	Standard dev. of buffer area covered by buildings

Table 3: The TFNs of the linguistic labels.

Label	Description	a	M	b
$\alpha_1 = Cv$	Optimum reliability	0.80	0.90	1.00
$\alpha_2 = V$	Good reliability	0.65	0.75	0.80
$\alpha_3 = Mv$	Sufficient reliability	0.55	0.60	0.65
$\alpha_4 = F$	Mediocre reliability	0.45	0.50	0.55
$\alpha_5 = \mathrm{Sc}$	Scanty reliability	0.35	0.40	0.45
$\alpha_6 = \mathrm{Bd}$	Bad reliability	0.20	0.30	0.35
$\alpha_7 = \mathrm{Nl}$	Null reliability	0.00	0.10	0.20

fuzzification process, we associate each isoreliable zone to a TFN as showed in Table 1.

A string is created for the parameter P_1 in the following form:

$$A = [O4]^{\alpha_6}[-]^{\alpha_5}[O5]^{\alpha_4}[-]^{\alpha_3}[O1, O3]^{\alpha_2}[O2]^{\alpha_1}, \quad (1)$$

where the symbol "–" indicates the absence of isoreliable zones to be associated with the corresponding TFN.

Now we suppose to create a partition in four TFNs of the domain of a second attribute P_2, obtaining the corresponding string B. The combination of the two strings A and B made by means of the new operation "Δ" is a string $C = (A\Delta B)$ (details are given in [6] and in Section 2.1) defined as

$$C = (A\Delta B) = [-]^{\gamma_8}[o_4]^{\gamma_7}[-]^{\gamma_6}[o_5]^{\gamma_5}[-]^{\gamma_4}[o1]^{\gamma_3}[o_3]^{\gamma_2}[o_2]^{\gamma_1}. \quad (2)$$

In this string new TFNs, labeled as $\gamma_1, \ldots, \gamma_8$, are obtained as well. Generally, if n (resp., m) is the number of TFNs for the string A (resp., B), then the string C contains $(n+m-1)$ TFNs. The strings obtained for each layer are recombined as in (2) for obtaining a final string used for evaluating the reliability of the whole set of layers. For instance, we can consider (see [4] for details) four aerophotogrammetric layers as given in Table 2.

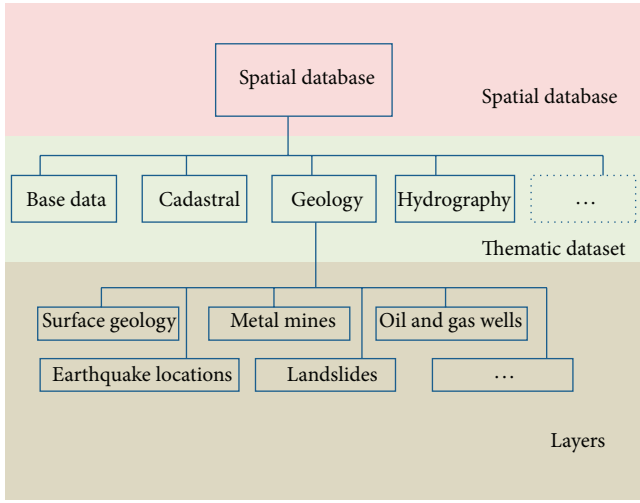

Figure 1: Spatial hierarchic database: three levels.

After calculating the strings for each layer, they are combined again by using operator (2) in order to obtain a final string. In this calculus a weight is associated with each layer, related to the role of that layer in the spatial database. For instance, a layer "buildings" can be more relevant with respect to the layer "infrastructures"; then the quality of the first layer affects the spatial database reliability more than the second one.

There is another question to be considered which consists in the fact that the weight associated with an attribute can change for different isoreliability zones. For instance, the quality of the dataset "spot elevation" affects the reliability of the spatial database in zones with strong slopes more than in smoothed zones. After assigning the weights, the final string is to be recalculated as showed in Section 2.3. The reliability index, to be assigned to each isoreliable zone, is given by the central value of the TFN in which that zone appears in the final string.

In our method we model a spatial database in three hierarchic levels: spatial database, thematic dataset, and layer. A spatial database is composed of several thematic datasets (e.g., geology, hydrology, etc.). Each thematic dataset contains more thematic layers (Figure 1).

Our method starts considering the layer level, assigning the single parameters to each layer. After determining the strings for each layer of a thematic dataset, they are combined with formula (2) for obtaining the final string of the thematic dataset; successively this string is recalculated by considering the weights assigned to each layer. After calculating the final strings of each thematic dataset, we combine these strings by using again formula (2) and we obtain the final string for the

whole spatial database. This string is to be recalculated by considering the weights assigned to each thematic dataset.

For each isoreliable zone, then we obtain the corresponding isoreliability index by taking the central value of the TFN related to that zone in the final string. We test our method considering the spatial database of the town of Cava de' Tirreni, near Salerno (Italy).

In Section 2 we present the algebraic structure given in [6]. Section 3 contains our method, Section 4 gives the results of our tests, and Section 5 is conclusions.

2. Definition of the Algebraic Structure

2.1. The Operations. We recall the main properties of the algebraic structure given in [6]. Let U be the universe of discourse and $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ an ordered n-tuple of linguistic labels, each composed from one or more linguistic modifiers and a variable, as, for example, "α_1 = False," "α_2 = More or Less Good," ...,"α_i = Good," "α_{i+1} = Very Good," ...,"α_n = Completely Good," and each represented by suitable TFNs denoted also by α_i, $i = 1, 2, \ldots, n$ (see, e.g., Table 3). Let A be a fuzzy attribute, that is, a map $A : U \to \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$, represented by a string of the following type:

$$A = [a_n]^{\alpha_n}[a_{n-1}]^{\alpha_{n-1}} \cdots [a_1]^{\alpha_1}, \tag{3}$$

where $a_i = A^{-1}(\alpha_i)$ is a subset of U, also called "class" in the sequel. If $A^{-1}(\alpha_i) = \emptyset$, then we write $a_i = [-]$. Let B be another fuzzy attribute represented by the following string:

$$B = [b_m]^{\beta_m}[b_{m-1}]^{\beta_{m-1}} \cdots [b_1]^{\beta_1}, \tag{4}$$

where the used symbols have a similar meaning to the above ones. In accordance with [6], we define the operation Δ between A and B by setting

$$C = (A \Delta B) = [c_{m+n-1}]^{\gamma_{m+n-1}}[c_{m+n-2}]^{\gamma_{m+n-2}} \cdots [c_1]^{\gamma_1}, \tag{5}$$

where, by assuming $n \geq m$ without loss of generality, the subsets $\{c_i\}$ are given from the following formulas for $i = 1, \ldots, m + n - 1$:

$$c_i = \begin{cases} \bigcup\limits_{j=1,\ldots,i} \left(a_{i-j+1} \bigcap b_j\right) & \text{if } 1 \leq i \leq m-1, \\ \bigcup\limits_{j=1,\ldots,m} \left(a_{i-j+1} \bigcap b_j\right) & \text{if } m \leq i \leq n-1, \\ \bigcup\limits_{j=i-n+1,\ldots,m} \left(a_{i-j+1} \bigcap b_j\right) & \text{if } n \leq i \leq m+n-1. \end{cases} \tag{6}$$

As suggested in [6], the subsets c_i can be calculated by using a simple rule based on the usual arithmetical multiplication. The TFNs γ_i, for $i = 1, \ldots, m+n-1$, are indeed given by

$$\gamma_i = \begin{cases} \dfrac{1}{k1 + k2} \cdot \sum\limits_{j=1}^{i} d2_j \cdot d1_{i-j+1} \cdot \left(k1 \cdot \alpha_{i-j+1} + k2 \cdot \beta_j\right) & \text{if } 1 \leq i \leq m-1, \\ \dfrac{1}{k1 + k2} \cdot \sum\limits_{j=1}^{m} d2_j \cdot d1_{i-j+1} \cdot \left(k1 \cdot \alpha_{i-j+1} + k2 \cdot \beta_j\right) & \text{if } m \leq i \leq n-1, \\ \dfrac{1}{k1 + k2} \cdot \sum\limits_{j=i-n+1}^{m} d2_j \cdot d1_{i-j+1} \cdot \left(k1 \cdot \alpha_{i-j+1} + k2 \cdot \beta_j\right) & \text{if } n \leq i \leq m+n-1 \end{cases} \tag{7}$$

with the above coefficients d_i, for $i = 1, \ldots, m+n-1$, defined by

$$d_i = \begin{cases} \sum\limits_{j=1}^{i} d2_j \cdot d1_{i-j+1} & \text{if } 1 \leq i \leq m-1, \\ \sum\limits_{j=1}^{m} d2_j \cdot d1_{i-j+1} & \text{if } m \leq i \leq n-1, \\ \sum\limits_{j=i-n+1}^{m} d2_j \cdot d1_{i-j+1} & \text{if } n \leq i \leq m+n-1. \end{cases} \tag{8}$$

The index $d1_i$ (resp., $d2_i$) represents the number of subsets $\{a_i\}$ (resp., $\{b_i\}$) of the string A (resp., B) involved in the operation of union performed to obtain the subsets $\{c_i\}$ of the resulting fuzzy attribute C, whereas the index $k1$ (resp., $k2$) stands for the total number of subsets $\{a_i\}$ of A (resp., $\{b_i\}$ of B) involved in the operation of intersection which gives the subsets $\{c_i\}$ of C.

2.2. The Weights of the Attributes. The first step, which precedes the above mentioned operations over the strings, consists in the determination of the weights of each attribute connected to a fixed zone because they can vary by changing zone. Strictly speaking, the above model implies the necessity to build a mean of the weights of the zones which have the same linguistic label in an attribute. This mean shall be the weight of that linguistic label, which in turn is multiplied for the middle point of the TFN, representing the same label, giving a number q, of which we consider the smallest integer contained in it, that is, INT (q). At the right of the same linguistic label, thus we create INT (q)-linguistic labels "approximated" with the procedure of Section 2.3. For example, we consider six zones $O1, \ldots, O6$ in which the fuzzy attribute A has received six values with the related weights $W1$ in accordance with Table 3. Then if $U = \{O1, O2, O3, O4, O5, O6\}$, then the fuzzy attribute A is represented by the following string:

$$\mathbf{A} = [O1, O3, O4]^{Cv}[O2]^{v}[O5]^{Mv}[O6]^{F} \tag{9}$$

and consider the linguistic label Cv. For simplicity, let us denote by W_{1i} the weight of the attribute A for the zones

Oi with $i = 1, 3, 4$. Then the mean value $W_{1,Cv}$ for Cv is equal to 2, to be multiplied for 1.0 (cf. Table 2) giving $N_{1,Cv} = \mathrm{INT}(W_{1,Cv} \cdot 1.0) = 2$ which represents the number of new linguistic labels, inserted at the right of Cv. Other new linguistic labels shall not be inserted at the right of the three remaining labels since we have, with evident meaning of the symbology, $W_{1,V} = W_{12} = 1$, $W_{1,Mv} = W_{15} = 1$, and $W_{1,F} = W_{16} = 1$ obtaining $N_{1,V} = \mathrm{INT}(W_{1,V} \cdot 0.7) = 0$, $N_{1,Mv} = \mathrm{INT}(W_{1,Mv} \cdot 0.6) = 0$, and $N_{1,F} = \mathrm{INT}(W_{1,F} \cdot 0.5) = 0$. Then we obtain the following finer string for the attribute **A**:

$$\mathbf{A} = [O1, O3, 04]^{Cv}[-]^{Cv,2}[-]^{Cv,1}[O2]^{V}[O5]^{Mv}[O6]^{F}. \quad (10)$$

This methodology gives the advantage to improve the position of the objects (in our case study, the isoreliable zones) in the set of the attributes, just bearing in mind the new linguistic labels with which the objects can be associated. The calculation of the membership functions for the TFNs, representing the new linguistic labels, is made in the following way.

Let β be the considered linguistic label present in the attribute **Ai** and let $N_{i,\beta}$ be the number of the new linguistic labels obtained with the above procedure. Let α be the linguistic label immediately following β in the linguistic labels of **Ai**. For every $t = 1, \ldots, N_{i,\beta}$, we put $a_{\beta,t} = a_\alpha + t^*(a_\beta - a_\alpha)/(N_{i,\beta} + 1)$ and similarly for $M_{\beta,t}$ and $b_{\beta,t}$. Then $[a_{\beta,t}, M_{\beta,t}, b_{\beta,t}]$ is the TFN representative of the linguistic label β, t.

2.3. Approximation of the Linguistic Labels. Some TFNs obtained in the final fuzzy attribute, after the successive composition of several strings, must be reconverted in linguistic labels, which can be approximated to known TFNs using the following procedure.

Let β be the TFN to be approximated and α, γ TFNs known (i.e., the meaning of their linguistic labels is known) such that $M_\alpha \leq M_\beta \leq M_\gamma$. By setting $d = M_\gamma - M_\alpha$ and if $M_\alpha \leq M_\beta \leq M_\alpha + d/10$, then we put $\beta = \alpha$; if $M_\alpha + d/10 < M_\beta \leq M_\alpha + 3d/10$, then we say β is "Next To α" and we write $\beta = \mathrm{NT}[\alpha]$; if $M_\alpha + 3d/10 < M_\beta \leq M_\alpha + 7d/10$, then we say β is "Included Between α and γ" and we write $\beta = \mathrm{IB}[\alpha, \gamma]$; if $M_\alpha + 7d/10 < M_\beta \leq M_\alpha + 9d/10$, then we say β is "Before To γ" and we write $\beta = \mathrm{BT}[\gamma]$; if $M_\alpha + 9d/10 < M_\beta \leq M_\gamma$, then we put $\beta = \gamma$. For instance, taking in account the TFNs of Table 3, let $\beta = \gamma_6$ as in Section 2.1. Since $M_V \leq M_\beta \leq M_{Cv}$ and $d = 0.30$, it is easily seen that $\beta = \mathrm{IB}[V, Cv]$.

We note that no matter of comparison between a_α, a_β, and a_γ and similarly for b_α, b_β, and b_γ is requested in this procedure.

3. Fuzzy Reliability for Spatial Databases

We model a *spatial database* in a three-level hierarchic structure as showed in Figure 1. The spatial database is composed of *thematic datasets* referred to as specific spatial domains. Each thematic dataset is composed of *layer*, that is, of georeferenced vectors or raster themes.

As in [4], after applying the algebraic structure on the strings corresponding to each parameter and recalculating the final string considering the weights assigned to the parameters, we reuse the algebraic structure operator applying it to the final strings associated with the layers of a thematic dataset and recalculating the final string obtained considering the weights associated with these layers. For obtaining the reliability index of the spatial database we apply the operator of the algebraic structure to the final strings associated with each thematic dataset and recalculate the obtained final string considering the weights assigned to each thematic dataset. Therefore we estimate the reliability of the spatial database in each isoreliable zone and apply the calculus on the algebraic structure as described in Section 2 in each level of our spatial database model. Below we describe the single steps that compose our method.

(1) The domain expert creates a partition of the area of study in isoreliable zones; each isoreliable zone is a geographical area, homogeneous in terms of data quality and environmental characteristics.

(2) For each layer the parameters are identified, that is, the observables that affect the quality of the layer, and assigned; for each isoreliable zone, the labels of the corresponding TFNs and the weights of each parameter are assigned as well. The expert creates a fuzzy partition in TFNs of the domain of each parameter.

(3) For each layer the operator of the algebraic structure [6] is applied on the parameters, obtaining a final string to be recalculated (as described in Section 2.2) by considering the weights assigned to the same parameters. We obtain the *index of reliability* of each layer; we call the map of this index *the reliability map* of the layer.

(4) In each thematic dataset its layers are identified as parameters; the string associated with a layer is given by the final string calculated for this layer. For each isoreliable zone the expert assigns the weights of each layer considering the impact of the layer on the quality of its thematic dataset.

(5) For each thematic dataset the operator of the algebraic structure [6] is applied on the related strings by obtaining a final string to be recalculated (as described in Section 2.2) by considering the weights assigned to the same layers. We obtain the *index of reliability* and the corresponding reliability map of the thematic dataset.

(6) Now we identify as parameters the thematic datasets of the spatial database; the string associated with a thematic dataset is given by the final string calculated for this thematic dataset. The expert assigns, for each isoreliable zone, the weights of each thematic dataset considering the impact of the layer on the quality of the spatial database.

(7) The operator of the algebraic structure [6] is applied on the string assigned to each thematic dataset, obtaining a final string to be recalculated (as described in Section 2.2) by considering the weights assigned to

Table 4: Values for W_1.

ID	A	W_1
$O1$	Cv	3
$O2$	V	1
$O3$	Cv	2
$O4$	Cv	1
$O5$	Mv	1
$O6$	F	1

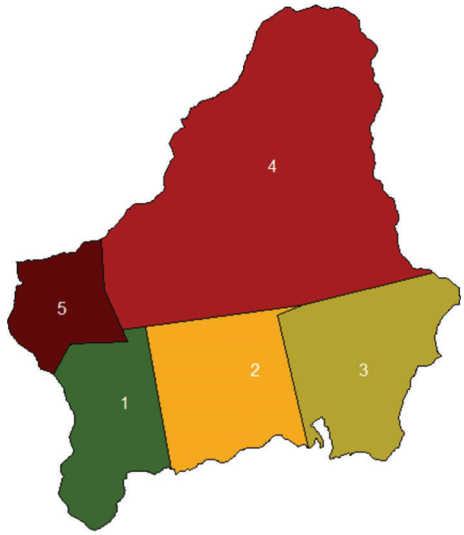

Figure 2: The five isoreliable zones of the area of study.

the same thematic datasets. We obtain the *index of reliability* and the related reliability map of the spatial database.

In Section 4 we present the results by applying our method on a spatial database based on the tool Fuzzy-SRA.

4. Test Results

In our tests the area of study is given by the town of Cava de' Tirreni (Italy). Considering the data quality and the environmental and climatic characteristics, the area of study is partitioned in five isoreliable zones as showed in Figure 2.

We consider the most significant thematic datasets and layers of the spatial database. In our test we consider 5 thematic datasets. In Table 5 we show the thematic datasets and the parameters chosen for each layer.

In the choice of the parameters, particular attention was focused on the absence of primary information connected to geographical entities (e.g., the height of a building or of a spot elevation's point).

Other characteristics that affect the quality of a layer consist of geometric and topological types of errors (isolated street lines or particles intersecting between them).

To simplify the calculus we create for each parameter a fuzzy partition in five TFNs, labeled as showed in Table 3. For brevity we show only the TFNs set for the parameter

Table 5: Layers and parameters of the spatial database.

Thematic dataset	Layer	Parameter
(1) Aerial photogrammetric data	(1.1) Streets	Density of isolated lines
		Density of unlabeled lines
	(1.2) Buildings	Density of intersecting polygons
		Density of polygons with uncoded height value
		Density of polygons with wrong height value
	(1.3) Spot elevations	Density of uncoded points
		Density of points with wrong elevation value
	(1.4) Contours	Density of uncoded lines
		Density of lines with wrong elevation value
(2) Cadastral data	(2.1) Terrain parcels	Density of intersecting polygons
		Density of uncoded polygons
	(2.2) Buildings parcels	Density of intersecting polygons
		Density of uncoded polygons
		Density of polygons not overlapping buildings
	(2.3) Land use	Density of intersecting polygons
		Density of unlabeled polygons
(3) Hydrology	(3.1) Rivers	Density of unlabeled lines
		Density of overlapping lines
	(3.2) Lakes	Density of intersecting polygons
		Density of unlabeled polygons
	(3.3) Aquifers	Density of polygons with wrong piezometric elevation value
(4) Geology	(4.1) Surface geology	Density of intersecting polygons
		Density of unlabeled polygons
	(4.2) Land slides	Density of intersecting polygons
		Density of unlabeled polygons
	(4.3) Alluvial zones	Density of intersecting polygons
		Density of unlabeled polygons

Table 5: Continued.

Thematic dataset	Layer	Parameter
(5) Infrastructural networks	(5.1) Road network	Density of overlapping lines
		Density of unlabeled arcs
	(5.2) Drainage system	Density of overlapping lines
		Density of unlabeled arcs
	(5.3) Water supply network	Density of overlapping lines
		Density of unlabeled arcs
	(5.4) Electricity grid	Density of overlapping lines
		Density of unlabeled arcs

Table 6: TFNs for the layer (1.1)—parameter "density of isolated lines."

Label	Description	a	M	b
Cv	Optimum reliability	0.00	0.005	0.01
V	Good reliability	0.01	0.015	0.02
Mv	Sufficient reliability	0.02	0.03	0.04
F	Mediocre reliability	0.04	0.05	0.06
Sc	Scanty reliability	0.06	0.08	0.10
Bd	Bad reliability	0.10	0.20	0.30
Nl	Null reliability	0.30	0.60	1.00

Table 7: TFNs for the Layer (2.1)—parameter "density of intersecting polygons."

Label	Description	a	M	b
Cv	Optimum reliability	0.00	0.005	0.01
V	Good reliability	0.01	0.02	0.03
Mv	Sufficient reliability	0.03	0.04	0.05
F	Mediocre reliability	0.05	0.07	0.09
Sc	Scanty reliability	0.09	0.12	0.15
Bd	Bad reliability	0.15	0.18	0.21
Nl	Null reliability	0.21	0.70	1.00

Table 8: Weights assigned for the layer (1.1)—parameter "density of isolated lines."

Zone	Weight
$O1$	4
$O2$	4
$O3$	4
$O4$	3
$O5$	3

"density of the isolated lines" of the layer 1.1 street and for the parameter "density of intersecting polygons of the layer 2.1—Terrain parcels."

For each isoreliable zone the weights of the parameters are assigned by an expert. For brevity, we show the weight assigned to the two previous parameters for each isoreliable zone.

Table 9: Weights assigned for the layer (2.1)—parameter "density of intersecting polygons."

Zone	Weight
$O1$	5
$O2$	5
$O3$	4
$O4$	4
$O5$	4

Table 10: Weights assigned to the layers of each thematic dataset for the isoreliable zone $O1$.

Thematic dataset	Layer	Weight
(1) Aerial photogrammetric data	(1.1) Streets	4
	(1.2) Buildings	4
	(1.3) Spot elevations	3
	(1.4) Contours	3
(2) Cadastral data	(2.1) Terrain parcels	4
	(2.2) Buildings parcels	4
	(2.3) Land use	3
(3) Hydrology	(3.1) Rivers	3
	(3.2) Lakes	3
	(3.3) Aquifers	2
(4) Geology	(4.1) Surface geology	4
	(4.2) Land slides	3
	(4.3) Alluvial zones	3
(5) Infrastructural networks	(5.1) Road network	4
	(5.2) Drainage system	3
	(5.3) Water supply network	3
	(5.4) Electricity grid	2

After combining the strings related to each parameter of a layer, we obtain a final string that is recalculated considering the weights assigned to the same parameters. By using this string we obtain the reliability map of the layer. For brevity, considering Tables 4, 6, 7, 8, 9, we show the isoreliability maps for the layers 1.1 (Figure 3) and 2.1 (Figure 4) with final string, respectively, given by

$$\mathbf{A} = [O1]^{Cv}[O2]^{\mathrm{IB}[V,Cv]}[O3]^{Mv}[O5]^{\mathrm{NT}[Mv]}[O4]^{F},$$
$$\mathbf{B} = [O1, O2]^{Cv}[O3]^{\mathrm{NT}[V]}[O4, O5]^{\mathrm{IB}[Mv,F]}. \tag{11}$$

In these maps we note that the two less reliable zones are $O4$ and $O5$; in fact, in these zones the data are imprecise. After obtaining the final string for each layer of a thematic dataset, we combine them for obtaining the final string for the thematic dataset. Then we recalculate this final string by considering the weights assigned to the layers for each isoreliable zone. In Table 10 we show the weights assigned to each layer for the isoreliable zone $O1$.

Table 11: Reliability values obtained for the five thematic datasets.

Isoreliable zone	Reliab. value 1	Reliab. value 2	Reliab. value 3	Reliab. value 4	Reliab. value 5
$O1$	0.90	0.87	0.85	0.85	0.87
$O2$	0.86	0.85	0.82	0.82	0.85
$O3$	0.67	0.74	0.74	0.71	0.71
$O4$	0.44	0.53	0.53	0.50	0.44
$O5$	0.47	0.57	0.57	0.53	0.47

Table 12: Weights assigned to the thematic datasets for the isoreliable zone $O1$.

Isoreliable zone	Weights of dataset 1	Weights of dataset 2	Weights of dataset 3	Weights of dataset 4	Weights of dataset 5
$O1$	5	5	3	4	3
$O2$	5	5	3	4	3
$O3$	5	5	3	4	3
$O4$	5	5	4	5	3
$O5$	5	5	4	5	3

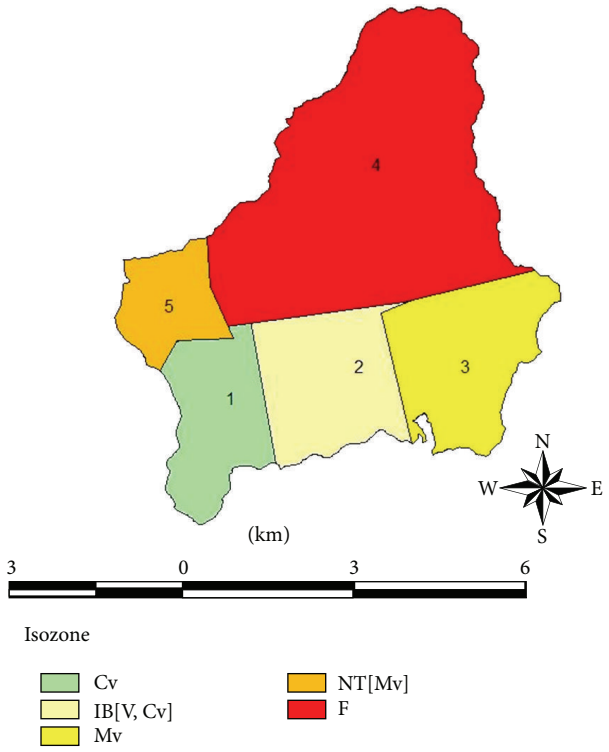

Figure 3: The reliability map for the layer 1.1—streets.

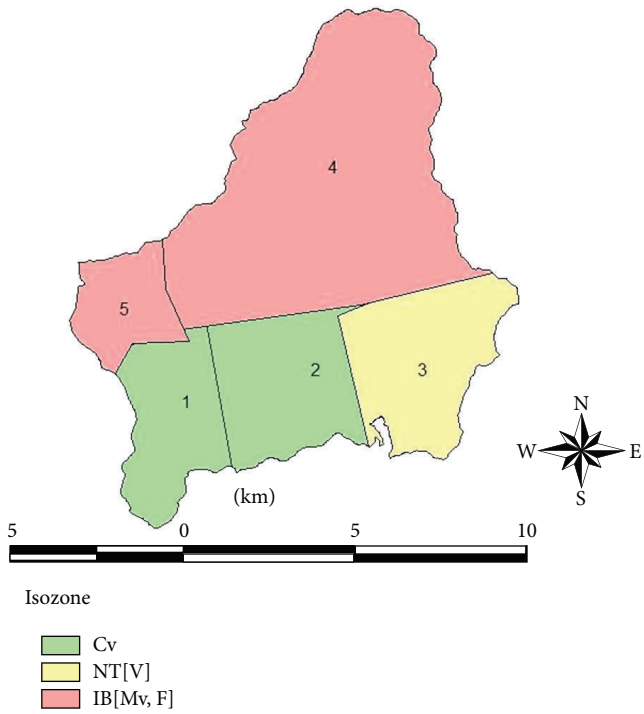

Figure 4: The reliability map for the layer 2.1—terrain parcels.

Figure 5 shows the thematic map of the thematic dataset "Aerial photogrammetric data." We consider as isoreliability values the central values of the TFN formed in the final string.

In Figure 6 we show the thematic map of the thematic dataset "Cadastral data."

The two reliability maps show that in the isoreliable zones $O4$ and $O5$ the quality of the data is poor. This result is confirmed for all the thematic datasets. In Table 11 we show the reliability values obtained for the five thematic datasets in each isoreliable zone.

After the calculation of the final strings for each thematic dataset, we combine them for obtaining the final string of the spatial dataset of Cava de' Tirreni (Italy). This final string is recalculated by considering the weights assigned to the thematic datasets for each isoreliable zone. The weights assigned for the five isoreliable zones are showed in Table 12.

The weights of the spatial datasets 3 and 4 for the isoreliable zones $O4$ and $O5$ are different with respect to the ones assigned for the isoreliable zones $O1$, $O2$, and $O3$. Indeed in the isoreliable zones $O4$ and $O5$ the surface terrains slopes are significant and there are many hydrographic characteristics.

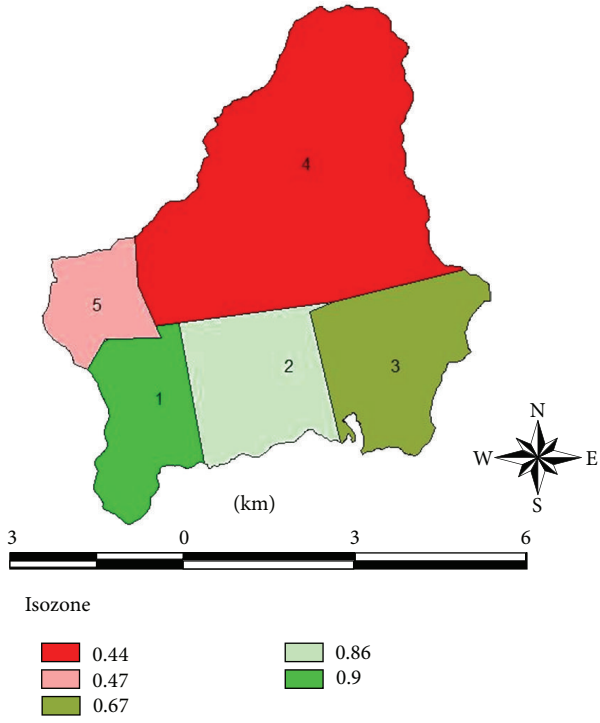

Figure 5: The reliability map for the thematic dataset "Aerial photogrammetric data."

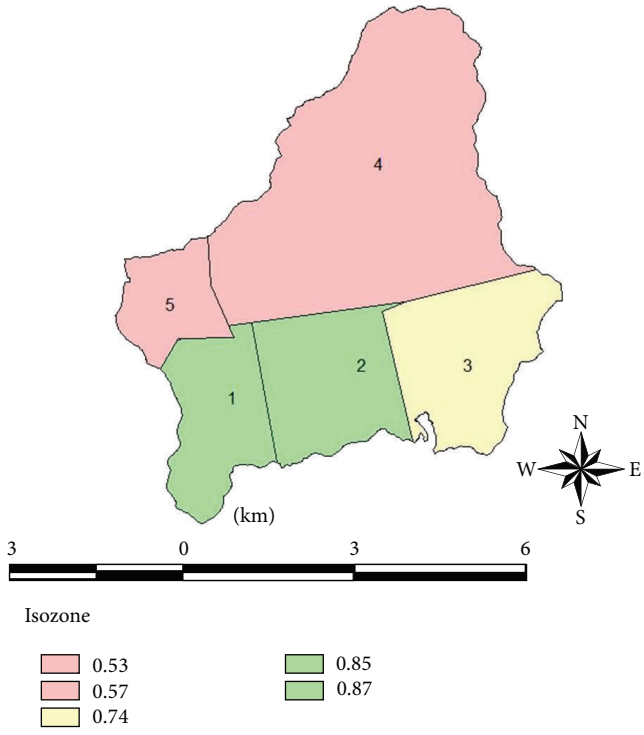

Figure 6: The reliability map for the thematic dataset "Cadastral data."

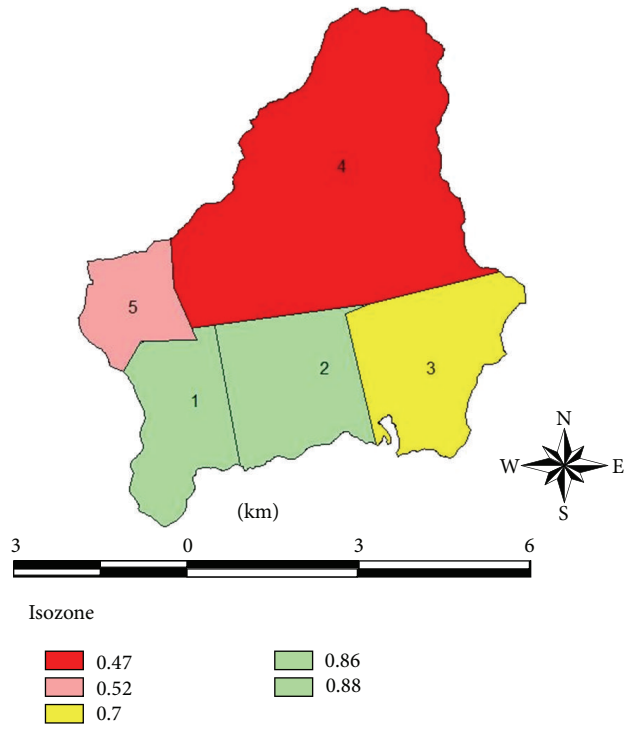

Figure 7: The reliability map for the spatial database of Cava de' Tirreni.

Finally in Figure 7 we show the reliability map for the spatial database.

The results in Figure 7 confirm the previous ones corresponding to the single thematic datasets. The reliability index is good in the isoreliable zones $O1$ and $O2$, fairly good in the isoreliable zone $O3$, and poor in the isoreliable zones $O4$ and $O5$.

5. Conclusions

To give an evaluation of the reliability of spatial database is a complex problem due to the lack of homogeneity of the spatial datasets and to the variation of the data quality on the area of study. Then the usage of a fuzzy logic approach is adequate for measuring the quality of spatial information. In this research we adopt the fuzzy algebraic structure [6] and the fuzzy reliability method applied in [4, 5] for evaluating the reliability of spatial datasets, in order to estimate the reliability of a whole spatial database.

We structure the spatial database in a three hierarchical levels, evaluating the reliability of the single layers, of the thematic datasets, and finally of the spatial database. We test our method on the spatial database of Cava de' Tirreni (Italy). An expert identifies the isoreliable zones and assigns the weights to each parameter, to the layers, and to the thematic datasets. We present the results obtained and the final reliability map of the spatial database on the area of study.

Acknowledgment

This work is performed in the context of the Project FARO 2010–2013 under the auspices of the "Polo delle Scienze e delle Tecnologie" dell'Università degli Studi di Napoli Federico II.

References

[1] A. Bardossy and L. Duckstein, *Fuzzy Rule-Based Modelling with Applications to Geophysical, Biological and Engineering Systems*, CRC Press, 1995.

[2] F. Di Martino and S. Sessa, "Spatial analysis with a tool GIS via systems of fuzzy relation equations," in *Computational Science and Its Applications: ICCSA 2011*, vol. 6783 of *Lecture Notes in Computer Science*, pp. 15–30, 2011.

[3] F. Di Martino and S. Sessa, "Spatial analysis and fuzzy relation equations," *Advances in Fuzzy Systems*, vol. 2011, Article ID 429498, 14 pages, 2011.

[4] F. Di Martino, V. Loia, S. Sessa, and M. Giordano, "An evaluation of the reliability of a GIS based on the fuzzy logic in a concrete case study," in *Fuzzy Modeling with Spatial Information for Geographic Problems*, F. E. Petry, V. B. Robinson, and M. A. Cobb, Eds., pp. 185–208, Springer, 2005.

[5] F. Di Martino, S. Sessa, and V. Loia, "A fuzzy-based tool for modelization and analysis of the vulnerability of aquifers: a case study," *International Journal of Approximate Reasoning*, vol. 38, no. 1, pp. 99–111, 2005.

[6] A. Gisolfi and V. Loia, "A complete, flexible fuzzy-based approach to the classification problem," *International Journal of Approximate Reasoning*, vol. 13, no. 3, pp. 151–183, 1995.

[7] F. Di Martino, V. Loia, and S. Sessa, "By fuzzy rules for applications to risk analysis in the Vesuvian area," *Contemporary Engineering Sciences*, vol. 4, no. 2, pp. 55–78, 2011.

Hindawi Publishing Corporation
Advances in Fuzzy Systems
Volume 2013, Article ID 876073, 7 pages
http://dx.doi.org/10.1155/2013/876073

Research Article

Hotspots Detection in Spatial Analysis via the Extended Gustafson-Kessel Algorithm

Ferdinando Di Martino[1] and Salvatore Sessa[1,2]

[1] *Dipartimento di Architettura, Università degli Studi di Napoli Federico II, Via Monteoliveto 3, 80134 Napoli, Italy*
[2] *Università degli Studi di Napoli Federico II, Centro Interdipartimentale per l'Analisi e la Progettazione Urbana Luigi Pisciotti, Via Toledo 402, 80134 Napoli, Italy*

Correspondence should be addressed to Ferdinando Di Martino; fdimarti@unina.it

Received 31 October 2013; Accepted 10 November 2013

Academic Editor: Sabrina Senatore

We show a new approach for detecting hotspots in spatial analysis based on the extended Gustafson-Kessel clustering method encapsulated in a Geographic Information System (GIS) tool. This algorithm gives (in the bidimensional case) ellipses as cluster prototypes to be considered as hotspots on the geographic map and we study their spatiotemporal evolution. The data consist of georeferenced patterns corresponding to positions of Taliban's attacks against civilians and soldiers in Afghanistan that happened during the period 2004–2010. We analyze the formation through time of new hotspots, the movement of the related centroids, the variation of the surface covered, the inclination angle, and the eccentricity of each hotspot.

1. Introduction

Hotspot detection is a known spatial clustering process in which it is necessary to detect spatial areas on which specific events thicken [1]; the patterns are the events georeferenced as points on the map; the features are the geographical coordinates (latitude and longitude) of any event. Hotspot detection is used in many disciplines, as in crime analysis [2–4], for analyzing where crimes occur with a certain frequency, in fire analysis [5] for studying the phenomenon of forest fires, and in disease analysis [6–9] for studying the localization and the focuses of diseases. Generally speaking, for detecting more accurately the geometrical shapes of hotspot areas algorithms based on density [10, 11] are used and they measure the spatial distribution of patterns on the area of study, but these algorithms have a high computational complexity.

In [5, 12, 13] a new hotspot detection method based on the extended fuzzy C-means algorithm (EFCM) [14, 15] was proposed, which is a variation of the famous fuzzy C-means (FCM) algorithm that detects cluster prototypes as hyperspheres. With respect to the FCM algorithm, the EFCM algorithm has the advantages of determining recursively the optimal number of clusters and being robust in the presence of noise and outliers. In [5, 12, 13] the EFCM is encapsulated in a GIS tool for detecting hotspots as circles displayed on the map. The pattern event dataset is partitioned according to the time of the event's detection, so each subset is corresponding to a specific time interval. The authors compare the hotspots obtained in two consecutive years by studying their intersection on the map. In this way it is possible to follow the evolution of a particular phenomenon studying how its incidence is shifting and spreading through time.

In this paper we present a new hotspot detection method based on the extended Gustafson-Kessel algorithm (EGK) [14, 15] for studying the spatiotemporal evolution of hotspots. Our aim is to improve the shape of the hotspots, maintaining a good computational complexity: indeed the EGK algorithm gives the cluster prototypes as hyperellipsoids and ellipses in the bidimensional case. The EGK algorithm is an extension of the Gustafson-Kessel (GK) algorithm [16] which we briefly present.

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset R^n$ be a dataset composed of N patterns $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T$, where x_{kj} is the kth

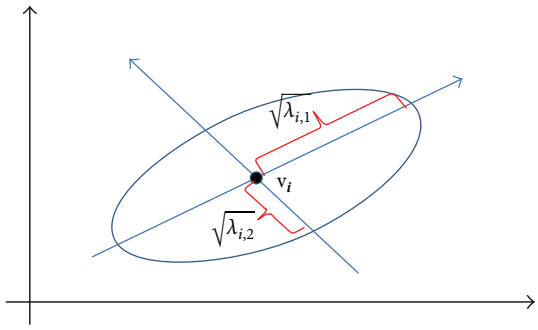

Figure 1: Example of ellipses cluster prototype using the GK algorithm.

component (feature) of the pattern $\mathbf{x}_j$. The GK algorithm minimizes the following objective function:

$$J(\mathbf{X}, \mathbf{U}, \mathbf{V}) = \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^m d_{ij}^2, \tag{1}$$

where C is the number of clusters fixed a priori, u_{ij} is the membership degree of the pattern $\mathbf{x}_j$ to the ith cluster ($i = 1, \ldots, C$), $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_C\} \subset R^n$ is the set of points given by the centroids of the C clusters, m is the fuzzifier parameter, and d_{ij} is the distance between $\mathbf{v}_i = (v_{1i}, v_{2i}, \ldots, v_{nj})^T$ and $\mathbf{x}_j$. The general form of this distance is given by

$$d_{ij} = \sqrt{\left(\mathbf{x}_j - \mathbf{v}_i\right)^T \mathbf{A}_i \left(\mathbf{x}_j - \mathbf{v}_j\right)}, \tag{2}$$

where $\mathbf{A}_i$ is the norm matrix, defined to be symmetric and positive. In the FCM algorithm $\mathbf{A}_i$ is equal to the identity matrix $\mathbf{I}$. In the GK algorithm the following Mahalanobis distance [17] is used:

$$d_{ij} = \sqrt{\left(\mathbf{x}_j - \mathbf{v}_i\right)^T \mathbf{P}_i^{-1} \left(\mathbf{x}_j - \mathbf{v}_j\right)}, \tag{3}$$

where $\mathbf{P}_i$ is the covariance matrix of the ith cluster given by

$$\mathbf{P}_i = \frac{\sum_{j=1}^{N} u_{ij}^m \left(\mathbf{x}_j - \mathbf{v}_i\right)\left(\mathbf{x}_j - \mathbf{v}_i\right)^T}{\sum_{j=1}^{N} u_{ij}^m}. \tag{4}$$

The covariance matrix $\mathbf{P}_i$ provides information about the shape and orientation of the cluster. The length of the kth axis of the hyperellipsoid is given by the root square of the kth eigenvalue λ_{ik} of $\mathbf{P}_i$. The directions of the axes of the hyperellipsoid are given by the directions of the eigenvectors of the matrix $\mathbf{P}_i$. In Figure 1 we show an example of ellipsoidal cluster prototype.

Using the Lagrange multipliers for minimizing objective function (1), we obtain the following solution for the centroids of each cluster prototype:

$$\mathbf{v}_i = \frac{\sum_{j=1}^{N} u_{ij}^m \mathbf{x}_j}{\sum_{j=1}^{N} u_{ij}^m}, \tag{5}$$

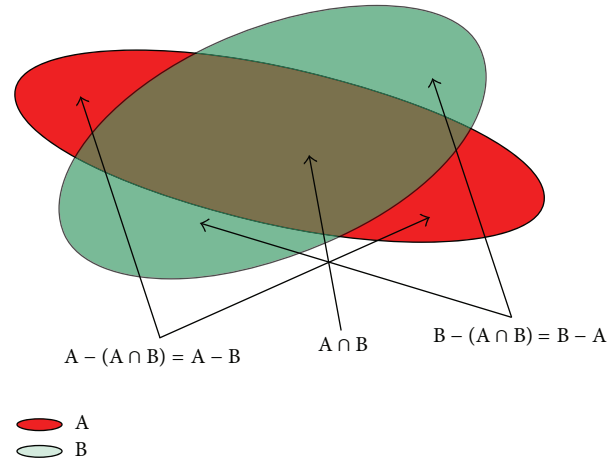

Figure 2: Intersection of two elliptical hotspots detected for events that happened in two consecutive periods.

where $i = 1, \ldots, C$ and u_{ij} are given by:

$$u_{ij} = \frac{1}{\left(\sum_{h=1}^{c} \left(d_{ij}^2 / d_{hj}^2\right)\right)^{2/(m-1)}}. \tag{6}$$

Initially the u_{ij}'s and the $\mathbf{v}_i$ are assigned randomly and updated in each iteration. If $U^{(l)} = (u_{ij}^{(l)})$ is the matrix $\mathbf{U}$ calculated at the lth step, the iterative process stops when

$$\left\|\mathbf{U}^{(l)} - \mathbf{U}^{(l-1)}\right\| = \max_{i,j} \left|u_{ij}^{(l)} - u_{ij}^{(l-1)}\right| < \varepsilon, \tag{7}$$

where $\varepsilon > 0$ is a prefixed parameter.

This algorithm is sensitive to the presence of outliers and noise and the number of cluster C is fixed a priori; as in the FCM algorithm, we need to use a validity index for determining an optimal value for the number of clusters C. In order to overcome these shortcomings, in [1, 16] the EGK algorithm is proposed which is a variation of the GK algorithm: there the optimal number of clusters is obtained during the iteration process. Furthermore, the EGK algorithm is robust with respect to the presence of noise and outliers.

In this paper we propose a new approach based on the EGK clustering method for detecting hotspots and studying their spatiotemporal evolution. Taking into consideration the bidimensional case, we obtain ellipses to be approximated as hotspot area better than the circular areas produced in the EFCM method.

Figure 2 shows an example of two intersecting elliptical hotspots, obtained as clusters detected by means of EGK method in two consecutive periods.

Figure 2 show three different regions:

(i) the area in which the hotspot A is not intersected by the hotspot B (corresponding to A − (A ∩ B) = A − B): this region can be considered as a set of geographical areas in which the prematurely detected event disappears successively;

(ii) the area of intersection A ∩ B: this area can be considered a geographical area in which the event persists in the course of time;

(iii) the area in which the hotspot B is not intersected by the hotspot A (corresponding to $B - (A \cap B) = B - A$): this region can be considered as a set of geographical areas in which the prematurely not detected event propagates successively.

We can study the spatiotemporal evolution of the hotspots by analyzing the interactions between elliptical hotspots detected for consecutive periods, by verifying the presence of clusters in areas in which clusters have not yet been detected previously and the disappearance of clusters in areas previously covered by hotspots.

In this research we present a method for studing the spatiotemporal evolution of hotspots areas of war in Afghanistan; we apply the EGK algorithm for comparing consecutive years' event datasets corresponding to positions of Taliban's attacks against civilian and soldiers. Each event corresponds to the geolocalization of the site where Taliban's attack happened as well.

We study the spatiotemporal evolution of the hotspots by analyzing the intersections of hotspots corresponding to two consecutive years, the displacement of the centroids, the increase or reduction of the hotspots areas, and the emergence of new hotspots.

In Section 2 we give an overview of the EGK algorithm. In Section 3 we present our method for studying the spatiotemporal evolution of hotspots in spatial analysis. In Section 4 we present the results of the spatiotemporal evolution of hotspots. Our conclusions are contained in Section 5.

2. The EGK Algorithm

In the EGK algorithm we consider clustering prototypes given by hyperellipsoids in the n-dimensional feature's space. The ith hyperellipsoidal cluster prototype V_i is characterized by a centroid $\mathbf{v}_i = (v_{1i}, \ldots, v_{ni})$ and a mean radius r_i and we say that x_j belongs to V_i if $d_{ij} \leq r_i$.

The radius r_i is obtained considering the covariance matrix $\mathbf{P}_i$ of the ith cluster, defined by (4), whose determinant gives the volume of the ith cluster. Since $\mathbf{P}_i$ is symmetric and positive, it can be decomposed in the form

$$\mathbf{P}_i = \mathbf{Q}_i \Lambda_i \mathbf{Q}_i^T, \tag{8}$$

where $\mathbf{Q}_i$ is an orthonormal matrix and $\lambda_i = (\lambda_{ij})$ is a diagonal matrix. The mean radius r_i is given as [15]

$$r_i = \frac{1}{n}\sqrt{\prod_{k=1}^{n} \lambda_{ik}^{1/n}} = \sqrt{\left(\det(\mathbf{P}_i)\right)^{1/n}}. \tag{9}$$

In the EGK algorithm the objective function to be minimized is

$$J(\mathbf{X}, \mathbf{U}, \mathbf{V}) = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}^{m}\left(d_{ij}^2 - r_i^2\right), \tag{10}$$

where d_{ij} is given by the Mahalanobis distance (3). The solutions obtained for the centroids $\mathbf{v}_i$ are given by formula (5) and the functions u_{ij} are obtained by

$$u_{ij} = \frac{1}{\sum_{h=1}^{C}\left(d_{ij}^2 w_{ij} / d_{hj}^2 w_{hj}\right)^{1/(m-1)}}, \tag{11}$$

where $w_{kj} = \max(0, 1 - (r_k^2/d_{kj}^2))$ for $k = 1, \ldots, C$. By setting $d_{kj}^2 w_{kj} = \max(0, d_{kj}^2 - r_k^2)$ and $\varphi_j = \text{card}\{k \in \{1, \ldots, C\} : d_{kj}^2 \leq r_k^2\}$ for any $j = 1, \ldots, N$, we have that formula (11) holds if $\varphi_j = 0$, while if $\varphi_j > 0$, one defines u_{ij} in the following way:

$$u_{ij} = \begin{cases} 0 & \text{if } d_{ij}^2 w_{ij} > 0 \\ \dfrac{1}{\varphi_j} & \text{if } d_{ij}^2 w_{ij} = 0. \end{cases} \tag{12}$$

Formula (12) produces the negative effect of diminishing the objective function (10) when a meaningful number of features are placed in a cluster; this effect can prevent the separation of the clusters. In order to solve this problem in [15], one starts with a small value r_i and by setting gradually $r_i := r_i + \beta^{(i)}/C^{(i)}$, where $C^{(i)}$ is the number of clusters at the ith iteration and $\beta^{(i)}$ is defined recursively as $\beta^{(0)} = 1$, $\beta^{(i)} = \min(C^{(i-1)}, \beta^{(i-1)})$ ($C^{(0)}$ is the initial number of clusters). By setting

$$I_{ik} = \frac{\sum_{j=1}^{N} \min\left(u_{ij}, u_{kj}\right)}{\sum_{j=1}^{N} u_{ij}}, \tag{13}$$

one defines the symmetric matrix $S = (S_{ik})$, where $S_{ik} = \max\{I_{ik}, I_{ki}\}$. If $S^{(l)}$ is the matrix S at the lth iteration ($l > 1$) and the threshold $\alpha^{(l)} = 1/(C^{(l)} - 1)$ is introduced, two indexes i^* and k^* are determined such that $S_{i^*k^*}^{(l)} = \max\{S_{ik}^{(l)} : i = 1, \ldots, N;\ k = 1, \ldots, C\} \geq \alpha^{(l)}$. Then these indices are merged by setting, for any $j = 1, \ldots, N$,

$$\begin{aligned} u_{i^*j}^{(l)} &:= u_{i^*j}^{(l)} + u_{k^*j}^{(l)}, \\ C^{(l)} &:= C^{(l-1)} + 1; \end{aligned} \tag{14}$$

thus, the k^*th row can be removed from the matrix $U^{(l)}$. In conclusion the following steps hold for the EGK algorithm.

(1) The user assigns initially $C^{(0)}$, $m > 0$ (usually $m = 2$), $\varepsilon > 0$, $S_{ik}^{(0)} = 0$, and $\beta^{(0)} = 1$.

(2) $u_{ij}^{(0)}$ are fixed randomly.

(3) v_i and r_i are calculated with formulae (5) and (9), respectively.

(4) u_{ij} are calculated with formulae (11) and (12).

(5) Determine i^* and k^* such that $S_{i^*k^*}^{(l)} = \max\{S_{ik}^{(l)} : i = 1, \ldots, N;\ k = 1, \ldots, C\}$.

(6) If $S_{i^*k^*}^{(l)} > \alpha^{(l)}$, then the i^*th and k^*th clusters are merged via formula (14) and the k^*th row is deleted from $U^{(l)}$.

(7) If formula (7) is satisfied, then the process stops otherwise go to (3) for the $(l+1)$th iteration.

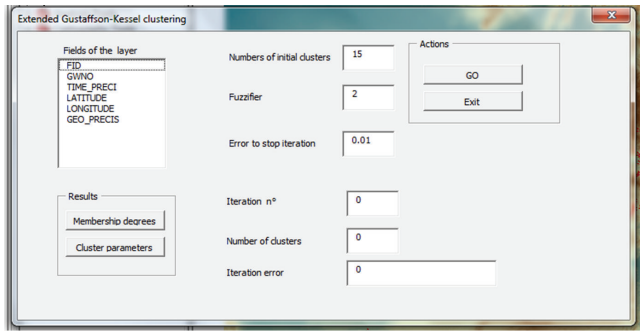

Figure 3: Mask created in the ESRI/ArcGIS tool for managing the EGK process.

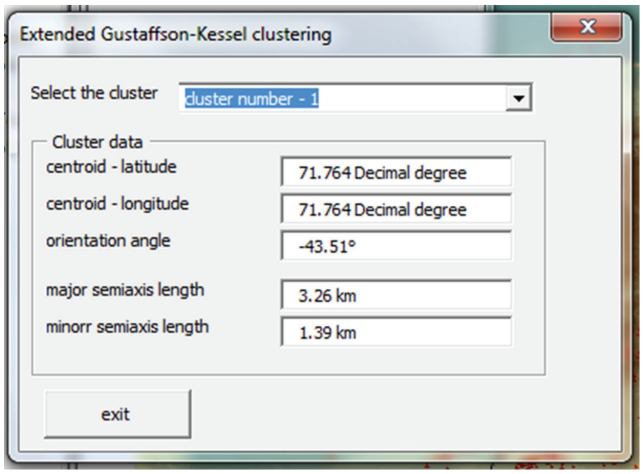

Figure 4: Mask created in the ESRI/ArcGIS tool for displaying the information of the detected cluster prototypes.

3. Hotspots Detection and Evolution in Military War

Each pattern is given by the event corresponding to a place in which an attack has occurred. The two features of the pattern are the geographic coordinates of this place.

We divide the event dataset into subsets corresponding to the events that occurred in a specific year or set of years. For each subset of events we apply the EGK algorithm to detect the final cluster prototypes.

The dataset is extracted from the URL http://www.acleddata.com/data/asia/; the data are the geolocalizations of Taliban's attacks in Afghanistan during the period 2004–2010. The EGK algorithm is encapsulated in the ESRI/ArcGIS tool. Figure 3 shows the mask used for setting the parameters and running the EGK algorithm.

We can set other numerical fields for adding other features to the geographical coordinates. Initially we set the initial number of clusters, the fuzzifier m (equal to 2 by default), and the error threshold for stopping the iterations (equal to 0.01 by default). At the end of the process we displayed on the form of the number of iterations, the final number of clusters, and the error calculated at the last iteration. The resultant clusters are shown as ellipses on the geographical map and can be saved in a new geographical layer.

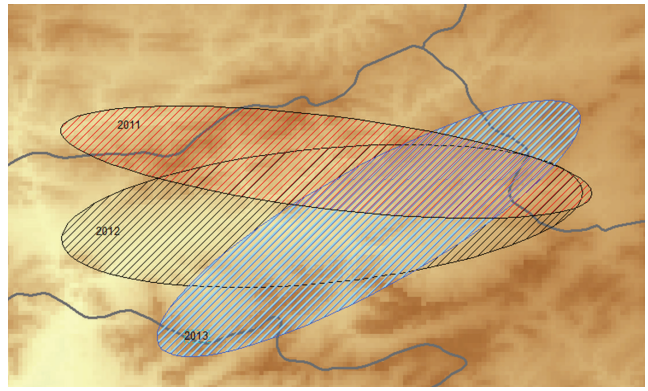

Figure 5: Spatiotemporal evolution of hotspots detected in three consecutive years.

Table 1: Results of the EGK applied to the event's subsets.

Year	Initial number of clusters	Final number of clusters	$\lvert U^{(l)} - U^{(l-1)} \rvert$	ε
2004–2006	15	7	0.79×10^{-2}	1×10^{-2}
2007	15	7	0.65×10^{-2}	1×10^{-2}
2008	15	7	0.68×10^{-2}	1×10^{-2}
2009	15	8	0.74×10^{-2}	1×10^{-2}
2010	15	8	0.81×10^{-2}	1×10^{-2}

In Figure 4 we show the mask used for displaying the information of each elliptical prototype detected: centroid's coordinates, length of each semiaxis, and orientation of the ellipses with respect to the horizontal plane on the geographical map.

The final process concerns the comparative analysis of the hotspots obtained by the final clusters resulting for each subset of events. Figure 5 shows an example of the display of hotspots obtained as final clusters corresponding to three consecutive years.

In order to assess the expansion and the displacement of any hotspot, we measure the area covered by each hotspot, the distance between the centroids of two intersecting hotspots detected in consecutive periods, the variation of the inclination angle, the eccentricity, and the length of both semiaxis.

4. Test Results

After partitioning the dataset in the five periods 2004–2006, 2007, 2008, 2009, and 2010, respectively, we apply the EGK algorithm for detecting the sequences of elliptical cluster prototypes. We fix $m = 2$, $C^{(0)} = 15$, and $\varepsilon = 1 \times 10^{-2}$. Table 1 shows the results obtained for each period.

We present the details relating to the comparison of the hotspots by considering the event data that occurred in the five periods. In Figures 6, 7, 8, 9, and 10 we show the hotspots detected.

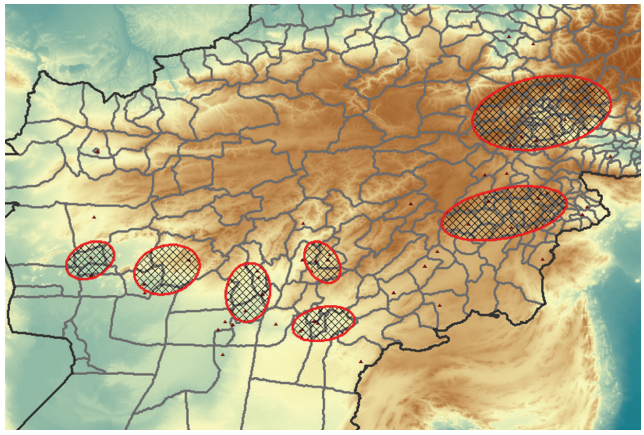

Figure 6: Taliban's attack events: years 2004–2006.

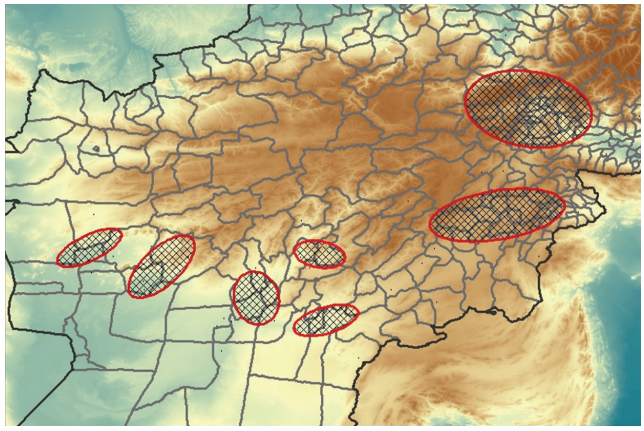

Figure 7: Taliban's attack events: year 2007.

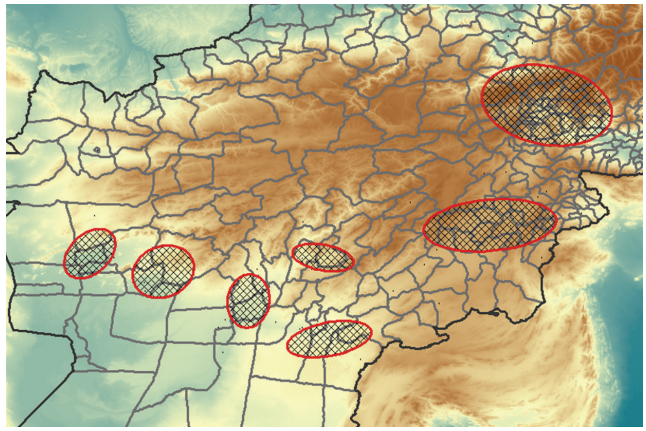

Figure 8: Taliban's attack events: year 2008.

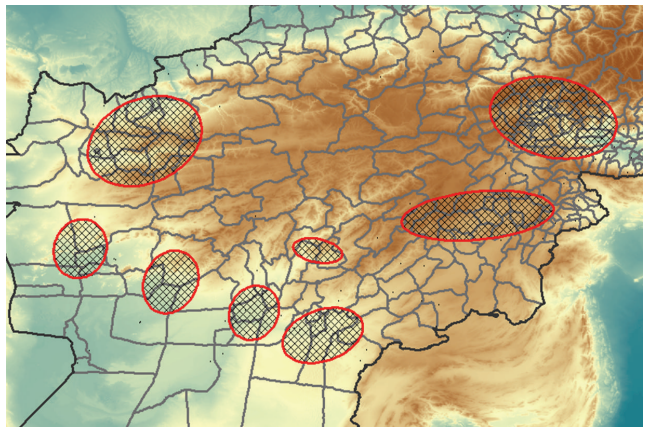

Figure 9: Taliban's attack events: year 2009.

Table 2: Areas of hotspots detected in 2009 and 2010.

Hotspot ID	Area hotspot 2009	Area hotspot 2010	Intersection area	% of Intersection area
1	18804.11	18856.72	14936.25	79.43%
2	13580.19	13583.15	10632.51	78.29%
3	1516.89	1538.23	1011.87	66.71%
4	7889.86	6665.24	5174.83	65.59%
5	4987.08	6086.51	3706.52	74.32%
6	6361.20	5691.76	4490.73	70.60%
7	5761.71	7866.39	5175.92	89.83%
8	18460.84	23471.45	16255.32	88.05%

By analyzing Figures 6–8 we can deduce that in the period 2004–2008 seven hotspot areas approximated as ellipses are present; in these periods each hotspot modified only slightly its angle, width, and position of the centroid. In the years 2009 and 2010 a new hotspot is detected in a region neighboring with Turkmenistan. In Figure 10 the hotspots obtained for two consecutive years 2009 and 2010 are overlapped as well. In blue (resp., red) we enumerate the hotspots corresponding to the year 2009 (resp., 2010); see Figure 11. The hotspots are labeled, and the hotspot number 8 is the new hotspot detected and coming from overlap of the related hotspots.

In Table 2 the first column shows the labels of each hotspot; the second and third columns show the area, in km^2, of the hotspot detected in 2009 and 2010, respectively. The fourth column (resp., fifth) shows the intersection area of the two hotspots (resp., the percentage of area of the hotspot detected in 2009 covered by the corresponding hotspot detected in 2010, that is, the ratio "intersection area/area hotspot detected in 2009").

The results in Table 2 show that over 65% of the area of each hotspot detected in 2009 is also covered by the corresponding hotspot detected in 2010. Another significant result is the increase of the area of the hotspot 8, which exceeds 2×10^4 km^2 in 2010. In Table 3 we show the eccentricity of each hotspot and the distance between the centroids of each hotspot detected in 2009 and the corresponding one detected in 2010.

The results show that the eccentricity increases significantly in 2010 for hotspots 4 and 6, whereas it decreases for hotspot 3; the eccentricity remains almost unchanged for the remaining hotspots in 2010. Another significant result is the

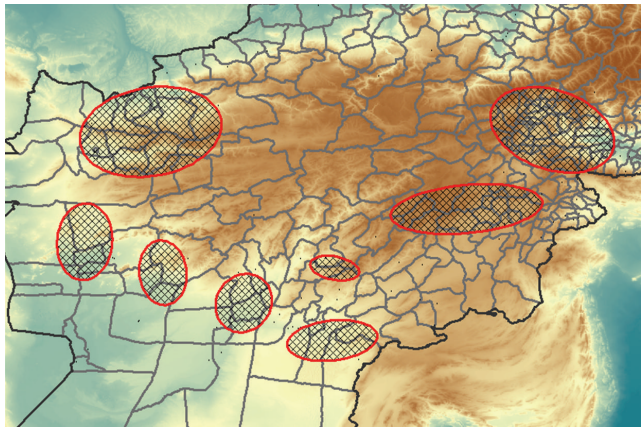

Figure 10: Taliban's attack events: year 2010.

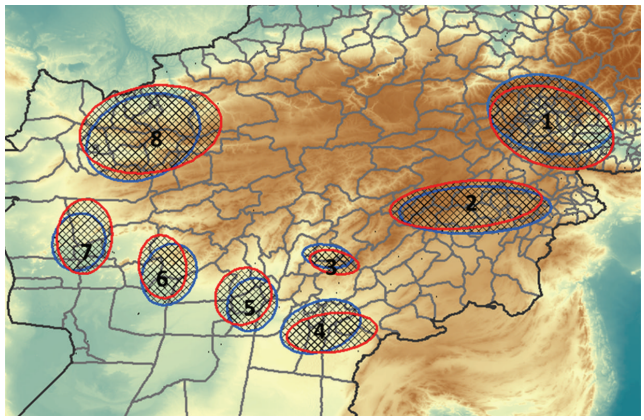

Figure 11: Hotspots detected for two consecutive periods: 2009 (blue) and 2010 (red).

Table 3: Eccentricity and centroid's distance between the hotspots detected in 2009 and 2010.

Hotspot ID	Eccentricity hotspot 2009	Eccentricity hotspot 2010	Centroids distance (km)
1	0.796	0.786	16.52
2	0.947	0.949	17.86
3	0.897	0.906	23.30
4	0.672	0.899	42.08
5	0.564	0.381	21.55
6	0.591	0.701	23.60
7	0.536	0.699	25.66
8	0.669	0.693	33.87

distance exceeding 40 km between the centroid of hotspot 4 detected in 2009 and the centroid of the corresponding hotspot detected in 2010.

5. Conclusions

We present a new approach for detecting hotspots in spatial analysis using the EGK clustering method encapsulated in a GIS tool. Similar to the EFCM algorithm, the EGK method is robust with respect to noise and outliers and we obtain the optimal number of the clusters iteratively during the process; furthermore, it has the advantage to detect hotspots of elongated shape. In our experiments we consider the site of Taliban's attacks in Afghanistan during the period 2004–2010. The spatial dataset is partitioned into subsets in order to study the evolution of the hotspots through time. We study the evolution of each hotspot in terms of movement of the centroids, surface covered, inclination, and eccentricity. The results show the formation, starting from 2009, of a new hotspot in the north-western zone neighboring with Turkmenistan. The results of the comparison of the hotspots detected in 2009 and 2010 show that this hotspot is increased with an extension of (about) 2×10^4 km^2.

Acknowledgment

This work is performed in the context of the project FARO 2010–2013 under the auspices of the "Polo delle Scienze e delle Tecnologie" of Università degli Studi di Napoli Federico II, Vapoli, Italy.

References

[1] T. H. Grubesic and A. T. Murray, "Detecting hotspots using cluster analysis and GIS," Annual Conference of CMRC, Dallas, 2001, http://www.ojp.usdoj.gov/ .

[2] S. P. Chainey, S. Reid, and N. Stuart, "When is a hotspot a hotspot? A procedure for creating statistically robust hotspot geographic maps of crime," in *Innovations in GIS 9: Socioeconomic Applications of Geographic Information Science*, D. Kidner, G. Higgs, and S. White, Eds., Taylor and Francis, London, UK, 2002.

[3] K. Harries, *Geographic Mapping Crime: Principle and Practice*, National Institute of Justice, Washington, DC, USA, 1999.

[4] A. T. Murray, I. McGuffog, J. S. Western, and P. Mullins, "Exploratory spatial data analysis techniques for examining urban crime," *British Journal of Criminology*, vol. 41, no. 2, pp. 309–329, 2001.

[5] F. Di Martino and S. Sessa, "The extended fuzzy C-means algorithm for hotspots in spatio-temporal GIS," *Expert Systems with Applications*, vol. 38, no. 9, pp. 11829–11836, 2011.

[6] M. R. Barillari, U. E. Barillari, F. Di Martino, R. Mele, I. Perfilieva, and S. Senatore, "Spatio-temporal hotspot analysis for exploring evolution of diseases: an application to oto-laryingo-pharingeal diseases," *Advances in Fuzzy Systems*, vol. 2013, Article ID 385974, 7 pages, 2013.

[7] R. M. Mullner, K. Chung, K. G. Croke, and E. K. Mensah, "Geographic information systems in public health and medicine," *Journal of Medical Systems*, vol. 28, no. 3, pp. 215–221, 2004.

[8] K. Polat, "Application of attribute weighting method based on clustering centers to discrimination of linearly non-separable medical datasets," *Journal of Medical Systems*, vol. 36, no. 4, pp. 2657–2673, 2012.

[9] C.-K. Wei, S. Su, and M.-C. Yang, "Application of data mining on the development of a disease distribution map of screened community residents of taipei county in Taiwan," *Journal of Medical Systems*, vol. 36, no. 3, pp. 2021–2027, 2012.

[10] I. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 773–780, 1989.

[11] R. Krishnapuram and J. Kim, "Clustering algorithms based on volume criteria," *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 2, pp. 228–236, 2000.

[12] F. Di Martino, V. Loia, and S. Sessa, "Extended fuzzy c-means clustering algo-rithm for hotspot events in spatial analysis," *International Journal of Hybrid Intelligent Systems*, vol. 4, pp. 1–14, 2007.

[13] F. Di Martino and S. Sessa, "Implementation of the extended fuzzy C-means algorithm in geographic information systems," *Journal of Uncertain Systems*, vol. 3, no. 4, pp. 298–306, 2009.

[14] U. Kaymak, R. Babuska, M. Setnes, H. B. Verbruggen, and H. M. van Nauta Lemke, "Methods for simplification of fuzzy models," in *Intelligent Hybrid Systems*, D. Ruan, Ed., pp. 91–108, Kluwer Academic, Dordrecht, The Netherlands, 1997.

[15] U. Kaymak and M. Setnes, "Fuzzy clustering with volume prototypes and adaptive cluster merging," *IEEE Transactions on Fuzzy Systems*, vol. 10, no. 6, pp. 705–712, 2002.

[16] D. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in *Procedings of the 17th IEEE Conf Decis Control Incl Symp Adapt Processes*, pp. 761–766, San Diego, Calif, USA, January 1979.

[17] R. Gnanadesikan and J. R. Kettenring, "Robust estimates, residuals, and outlier detection with multiresponse data," *Biometrics*, vol. 28, pp. 81–124, 1972.

Hindawi Publishing Corporation
Advances in Fuzzy Systems
Volume 2013, Article ID 506406, 7 pages
http://dx.doi.org/10.1155/2013/506406

Research Article

Usage of Fuzzy Spatial Theory for Modelling of Terrain Passability

Alois Hofmann,[1] Sarka Hoskova-Mayerova,[2] and Vaclav Talhofer[1]

[1] *Department of Military Geography and Meteorology, Faculty of Military Technology, University of Defence, Kounicova 65, 66210 Brno, Czech Republic*
[2] *Department of Mathematics and Physics, Faculty of Military Technology, University of Defence, Kounicova 65, 662 10 Brno, Czech Republic*

Correspondence should be addressed to Sarka Hoskova-Mayerova; sarka.hoskova@seznam.cz

Received 14 September 2013; Accepted 10 October 2013

Academic Editor: Ferdinando Di Martino

Geographic support of decision-making processes is based on various geographic products, usually in digital form, which come from various foundations and sources. Each product can be characterized by its quality or by its utility value for the given type of task or group of tasks, for which the product is used. They also usually have different characteristics and thus can very significantly influence the resulting analytical material. The aim of the paper is to contribute to the solution of the question of how it is possible to work with diverse spatial geographic information so that the user has an idea about the resulting product. The concept of fuzzy sets is used for representation of classes, whose boundaries are not clearly (not sharply) set, namely, the fuzzy approach in overlaying operations realized in ESRI ArcGIS environment. The paper is based on a research project which is being solved at the Faculty of Military Technologies of the University of Defence. The research deals with the influence of geographic and climatic factors on the activity of armed forces and the Integrated Rescue System.

1. Introduction

Geographic support for decision-making processes is based on various geographic products, usually in digital form, which come from various foundations and sources. Each product can be characterized by its quality [1, 2] or by its utility value for the given type of task or group of tasks, for which the product is used [3, 4]. In both cases, among others also positional and thematic accuracy are evaluated either as an exactly given value, for example, mean square position error, probable error, and so forth, or as a level of fulfilment of user's requirements expressed in percentage [5]. Decision-making processes are based on multicriteria decision-making, in which many factors are involved [6]. By a suitable combination of various factors, analytical products are created. They are the base for answering questions such as "What happens if…?" That is the reason why various foundations are used for geographic support. They also usually have different characteristics and thus can very significantly influence the resulting analytical material. The aim of the paper is to contribute to the solution of the question of how it is possible to work with diverse spatial geographic information so that the user has an idea about the resulting product. The paper is based on a research project which is being solved at the Faculty of Military Technologies of the University of Defence. The research deals with the influence of geographic and climatic factors on the activity of armed forces and the Integrated Rescue System.

Within the work on solution of the project, the problem solving team continued with development of models of influences of geographical and climatic factors on terrain passability. In the evaluated period, the team focused on development of the theory of models, on ways of visualization of vagueness of results caused by inhomogeneity of the support data, combining their characteristics within the performed analyses, and so forth. For these purposes, fuzzy logic was used.

2. Vagueness of Spatial Information

Standardly used types and procedures of geoprocedural analyses suppose that the used data are created with explicit, sharp boundaries that restrict the position of the individual objects.

The technical quality of position setting of such boundaries is usually given by the above-mentioned criteria, such as mean square positional error of its determination. The technical quality is based on measurement methods or digitalization of foundation.

Boundaries of many geographical elements are, however, only a result of human perception and not of a real matter of things. Even in case of existence of a real discrete boundary, the boundary line may be inaccurate due to vagueness of data or their interpretation. Vegetation or soil types are typical examples of geographical elements where there are no clear natural boundaries in space. Traditional classifications fail completely here. However, spatial units are usually represented by sharp boundaries. Brown and Heuvelink [7] suggest 2 types of uncertainty that have a special importance in GIS—*thematic* and *spatial vagueness*. In the first case we are not able to confirm occurrence of the given topic in the given place, and in the second case it is an inability to find an exact placement of the given topics. The formal conceptual model of uncertainty is useful to create for effective GIS utilization in the decision-making process [8]. Several authors tried to classify uncertainty [9–12], but the most used classification can be found in the standard for spatial data transfer [13].

As people both reason and make decisions with uncertain geospatial data every day, it is important to understand the complexity of uncertainty, how it propagates through each dataset, and how to best visualize uncertainty to support reasoning and decision-making [14–18].

Spatial and thematic data should not be evaluated independently. The way to solve these problems is to use the theory of "soft classifications," among which also *"fuzzy" approaches* belong. Fuzzy logic is quite frequently used for uncertainty of GIS data expression in the last years and is based on fuzzy theory [19–21].

Fuzzy sets together with the theory of fuzzy logic offer a scope for processing of predicates, whose level of probability is given in degrees ("true to certain degrees") and uncertainty is also expressed gradually. The concept of fuzzy sets deals with representation of classes, whose boundaries are not clearly (not sharply) set. When sharp boundaries separating the set from the surroundings are missing, a problem of unambiguous setting of an element belonging to a set and to its complement appears [22, 23].

Fuzzy files are then files or classes that do not have sharp limitation. With spatial data it means that at considered places the transition between the membership and nonmembership in a file is gradual. The fuzzy file then can be characterized by fuzzy levels of membership in an interval from 0.0 to 1.0 which expresses a gradual growth of membership from nonmembership up to the full membership. It can be defined with the help of the membership function.

In the environment of GIS three basic types of geoelements are usually defined: points, lines and areas (polygons). When we use linesm and areas we sometimes ask a question of how to delimit boundaries of the given geoelement. If there is an area layer which captures ecological stability of the given area, then there are only two possibilities of how to express stability: stable or unstable. This classification is very difficult and it depends on the person that decides and on the concrete area.

One of the basic characteristics that can be defined when creating and saving geographical objects is *topology*. Topological relations characterize the relative placement of two spatial objects with respect to their mutual position—for example, if they touch, overlay, or contain one another. In GIS they are important especially for a definition of spatial questions and selections and they play an important role when language SQL is used. In case of fuzzy spatial objects, however, traditional topologic predicates fail and their fuzzy variants come into consideration; they are able to answer inquiries such as the following.

(i) Do areas A and B overlay at least a bit?

(ii) Does area A partly contain area B?

(iii) Which areas are partly inside area B?

The fact that belonging of an element to a fuzzy topological predicate is expressed by a set [0, 1], however, complicates its direct usage in language SQL and thus possible spatial inquiries.

3. Fuzzy Overlay

A concrete possibility of usage of the fuzzy approach is the application in overlaying operations that can be realized in the environment of ArcGIS [24]. For this type of operations it is nowadays possible to use tools of map algebra as well as already implemented tools in the extension of Spatial Analyst. Fuzzy logic in the so-called "Overlay Analysis" is based on two fundamental steps.

The first one is the so-called "*fuzzification*" *or "fuzzy membership,*" in other words the process of implementing values into fuzzy sets, and the second step is the actual analysis-overlay of these sets. It is possible to use several types of fuzzification functions, from a linear to Gaussian function, according to the type of value distribution.

Relations among created fuzzy sets are then *analysed* by means of *fuzzy overlaying operations*, such as FuzzyAnd and FuzzyOr. The detailed description can be found for example, in the help file of ArcGIS Help.

Practical approaches of realization of usage of the "fuzzy" approach in analysis of geospatial data were the subject of development of procedures of finding an optimal route with the help of map algebra, which was suggested as one of the results of the mentioned project. It was testing of data that showed positional vagueness.

4. Terrain Passability

A very frequent task in the decision-making process in armed forces is to evaluate the possibility of vehicle movement in a terrain. This task is usually called Cross-Country Mobility

(CCM) in military language. The main goal of CCM is to evaluate the impact of geographic conditions on movement of vehicles in terrain [25, 26]. For the purpose of classification and qualification of geographic factors of CCM, it is necessary to determine

(i) particular degrees of CCM,

(ii) typology of terrain properties evaluated by the kind of vehicles used,

(iii) geographic factors and features with significant impact on CCM.

As a result of the geographic factors impact evaluation we get three CCM degrees: passable terrain, passable terrain with restrictions, or impassable terrain.

The impact of a geographic factor can be evaluated as a *coefficient of deceleration "C_i"* on the scale of 0 to 1. The coefficient of deceleration shows the real (simulated) speed of vehicle v in the landscape in the confrontation with the maximum speed of a given vehicle $v_{\max}$. The impact of all n geographic factors can be expressed by the following formula:

$$v = v_{\max} \prod_{i=1}^{n} C_i, \quad n = 1, \ldots, \mathbb{N}. \tag{1}$$

The main coefficients of deceleration are listed in Table 1.

In case of searching for an optimal route within the project, a procedure of fuzzification for calculation of coefficient of deceleration C_3 (influence of soil and ground cover) was looked for. The initial data layer was a raster layer of coefficient C_3, where there are 3 values for passable (1), passable with difficulty (0.5), and impassable terrain (0). From the point of view of the input data—that is, area of soil types and kinds—a presumption was introduced that the boundary of passability will not be sharp according to the original data basis but it will change its value in the distance of 100 m to both sides. Based on this presumption, fuzzy sets were created with the use of several raster techniques and their overall influence was analysed.

In the next step, possibilities of analysis of several resulted sets were tested by their mutual interaction (overalls) through various methods. The solution of the problem, however, was not finished in such a way so that it was possible to prove the change of course of the found route due to vagueness. The authors suppose that the stated results will be further developed in the years to come when solving the project and their successive transfer to models with vagueness consideration. The way of solution is discussed in the following paragraphs.

5. Development of Creation of CCM Models Using Fuzzy Logic

Six basic models were created for searching of the way. With the help of these models it is possible to create a so-called Cost Map, which is a raster file that is the basic input information for creation of the file of the searched route. The cost map is made by application of overlaying operations or in this case with the use of the so-called map algebra that provides tools for working with raster files. To set the overall coefficient of deceleration when moving through terrain, relations stated in the elaborated methodology were used.

Table 1: Main coefficients of deceleration.

Basic coefficient	Geographic signification and impact
C_1	Terrain relief (gradient of terrain relief and microrelief shapes)
C_2	Vegetation cover
C_3	Soils and soil cover
C_4	Weather and climate
C_5	Hydrology
C_6	Built-up area
C_7	Road network

Models for calculation of individual coefficients were compiled with the help of basic operations with raster data by means of implemented tools of the so-called Spatial Analyst, which is an extension of ArcGIS system, as well as using a tool of the so-called map algebra, that is, a set of operators and functions for work with raster data.

Raster layers were used as input data. The layers were created by calculation (e.g., a high raster model) or by a conversion of vector data according to the appropriate attributes. This data base was created as explicit data with clearly defined objects and their boundaries.

Another step when solving CCM problem is the introduction of some uncertainty causing a greater activity in decision-making about the use of the gained results. For each coefficient of deceleration, a new process model was made, in which principles of fuzzy logic were applied. For solution of the individual models, various approaches had to be used with respect to the character of input data and result that shall be reached.

5.1. Creation of Models for the Individual Coefficients. In the following text are briefly discussed procedures of calculation of the individual coefficients of deceleration that were realized in the environment of ArcGIS.

5.1.1. Coefficient of Deceleration due to Elevation of Terrain. Layers of relief and objects of microrelief are taken as source data. The compiled model is created from a layer of relief that was calculated with the help of raster interpolation and layers of point, line, and polygons.

In case of raster relief, the value of height is determined for each individual pixel of high area. In this case for its fuzzification we used data about the height range of the height model and setting of boundary for determination of impassability for the given means of transport.

For other layers the procedure is based on several substeps. The first step is the conversion of the layer in a raster and its editing. Data about height and depth of an object were used to convert height objects to a raster format. During the conversion of a vector into a raster it happens that the individual objects are converted into a set of pixels with a value of an attribute, according to which the conversion

is realized (e.g., hgt, which is a height of an object). All surrounding pixels are marked with attribute NoData. These pixels would not enter into the calculation of the cost map not only of this but also of all other overlaying raster layers. The editing is testing of individual pixels for attribute NoData (by function IsNull) and assigning value 0 to all found pixels. Thus it is secured that all pixels outside height objects have a zero value which is later used when creating the "Cost Map". This procedure is applied also for other input layers.

The second step is a conversion of values "integer" raster to values in a real domain so that the operation of "fuzzification" could be initiated, that is, implementing of values of the raster into a "fuzzy set" with the help of one of the "fuzzification" functions. The input raster was in these cases divided by the value 10.0 so that a raster with real values was created. For verification of the "fuzzy" approach to this problem, only linear function was used so far in all phases of calculation.

The third step of the solution is the use of "fuzzy" sets for application of vagueness in the position of the given types of objects. This problem is solved with the help of calculation of "Euclidean" distance for each type of objects. Placing values of distance into "fuzzy" sets uses once again a linear function with setting of various values of maximum distance. It depends on the values of four classes of accuracy that the database creator gives [27]:

(i) <0.5 meter for geodetic points,

(ii) <3 metres for stable objects,

(iii) <10 metres for other objects,

(iv) <20 metres for unstable objects.

Apart from these values it is necessary to consider also the size of a pixel that is used for conversion of vector drawing to a raster. The above stated values make sense in case the size of a pixel is maximally 1 meter.

For comparison, cost maps were created with a pixel value of 5 metres and fuzzification of distance was in one case chosen the same as the size of a pixel (which actually represents a sharp boundary and thus it does not get blurred) and in another case the distances are chosen mostly as 100 m besides calculation of factors C_6 and C_7 where the value was reduced to 30 m with respect to the character of objects.

The last step for most calculated factors is a combination of the gained results with the help of the so-called "Fuzzy Overlay," that is, weighted "fuzzy" overlay of the individual raster layers (within calculation of factor C_1).

Overlay can be realized with the use of various logical operations, in the case of logical sum "OR" that made sure the highest value from the overlaying rasters was written into the resulting pixel; in most cases it was value "1" representing basically impassable terrain (see Figures 1 and 2 in detail).

5.1.2. Coefficient of Deceleration due to Vegetation. In case of vegetation, the type of vegetation is significant information. If it is vegetation with grown trees, information about trunk cross-section and spacing between trees are taken as parameters. For fuzzification, such a procedure was chosen that stems from classification of vegetation according to a value of these parameters and introduction of geometric vagueness of boundaries of in this way classified vegetation. The result is a layer, in which appropriate values are assigned to areas and their boundaries are blurred with the help of a distance function and their values change depending on the type of fuzzification method (Figures 3 and 4).

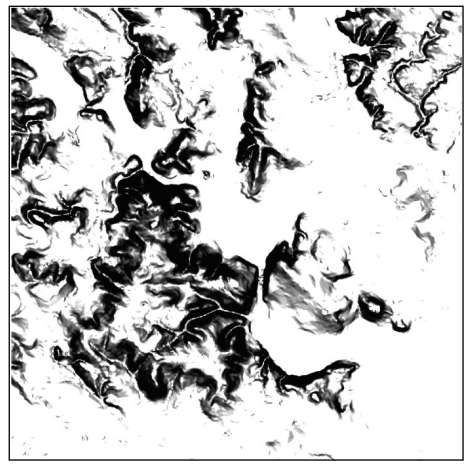

Figure 1: Visualization of calculation results of coefficient C_1.

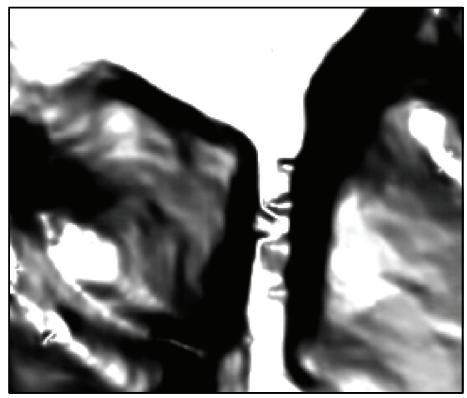

Figure 2: Visualization of calculation results of coefficient C_1-detail.

5.1.3. Coefficient of Deceleration due to Soils. For calculating this coefficient, attributes soil type, soil kind, and matrix were used. Using logical operators helped to determine passability of soils in four values (1, 2, 3, and 4). For fuzzification a procedure that is based on these values was used together with application of the general steps stated before, with the difference in vegetation lying in the method of creation of vague boundaries of the individual classes of soil polygons. In this case, these classes had to be transformed into independent layers for which distance rasters for blurring of their boundaries were calculated. Unlike the previous values of accuracy that were applied for all other factors of deceleration, boundaries for blurring with soils were chosen to be 100 m with respect to the character of this geographical element. The calculated raster entered into other steps of

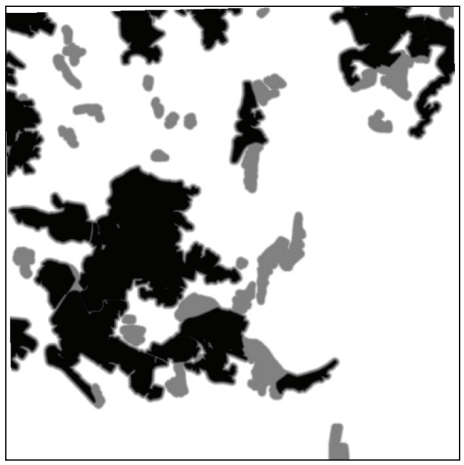

Figure 3: Visualization of calculation results of coefficient C_2.

Figure 4: Visualization of calculation results of coefficient C_2-detail.

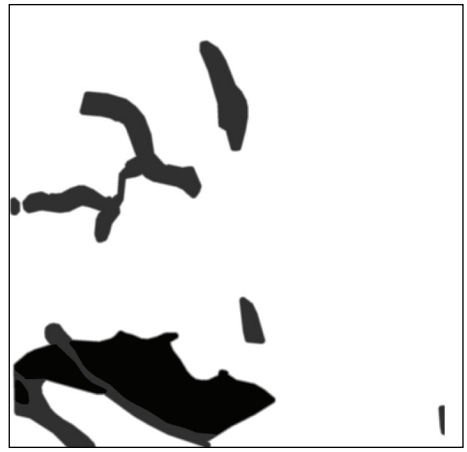

Figure 5: Visualization of calculation results of coefficient C_3.

Figure 6: Visualization of calculation results of coefficient C_3-detail.

calculation depending on the set value, especially by keeping of the already calculated fuzzified value for various levels of passability. This is realized with the help of multiplication process of the individual fuzzy distance raster by the fuzzy raster of passability.

The result is a raster with different values of passability and blurred boundaries of polygons (see Figures 5 and 6).

5.1.4. Coefficient of Deceleration due to Water Courses and Areas of Water. The calculation of this coefficient is rather complicated with respect to the fact that there should be quite a lot of input layers of various characters entering the calculation. When modelling, a reduced scale of input layers was used with regards to the fact that appropriate attributes were not available (e.g., character of bottom or banks). For creation of the model, a wider range of processes was used and during the calculation also more semiresults occurred that entered into the calculation of the final overlaying fuzzy operation.

The evaluation itself of water objects was different with respect to its modelling (lines object for streams and narrow rivers polygons for wide rivers, lakes, and ponds, etc.). For line water, attributes such as depth and width of the course were taken from the value of pixels creating a raster line, for areal; however, the width is given by the total area of pixels creating the area of water. Depth is then written into each pixel of the object in the same value.

For distance calculations, distance rasters for flowing and slack water were calculated separately. The result of the calculation can be visualized similarly.

5.1.5. Coefficient of Deceleration due to Built-Up Area. The solution of calculation of the model of a built-up area is divided into two parts. The first part is based on a presumption that all buildings are impassable and only the principle of vagueness of boundaries or position of individual objects in a layer of building is applied. The second part works with the possibility of different passability through a built-up area. It is based on attributes from a layer of block built-up area and it counts with different passability for various types of a built-up area.

5.1.6. Coefficient of Deceleration on Communications. The problem of calculation of coefficient of deceleration for communications is made difficult by the fact that for most motorized means of transport communications are well

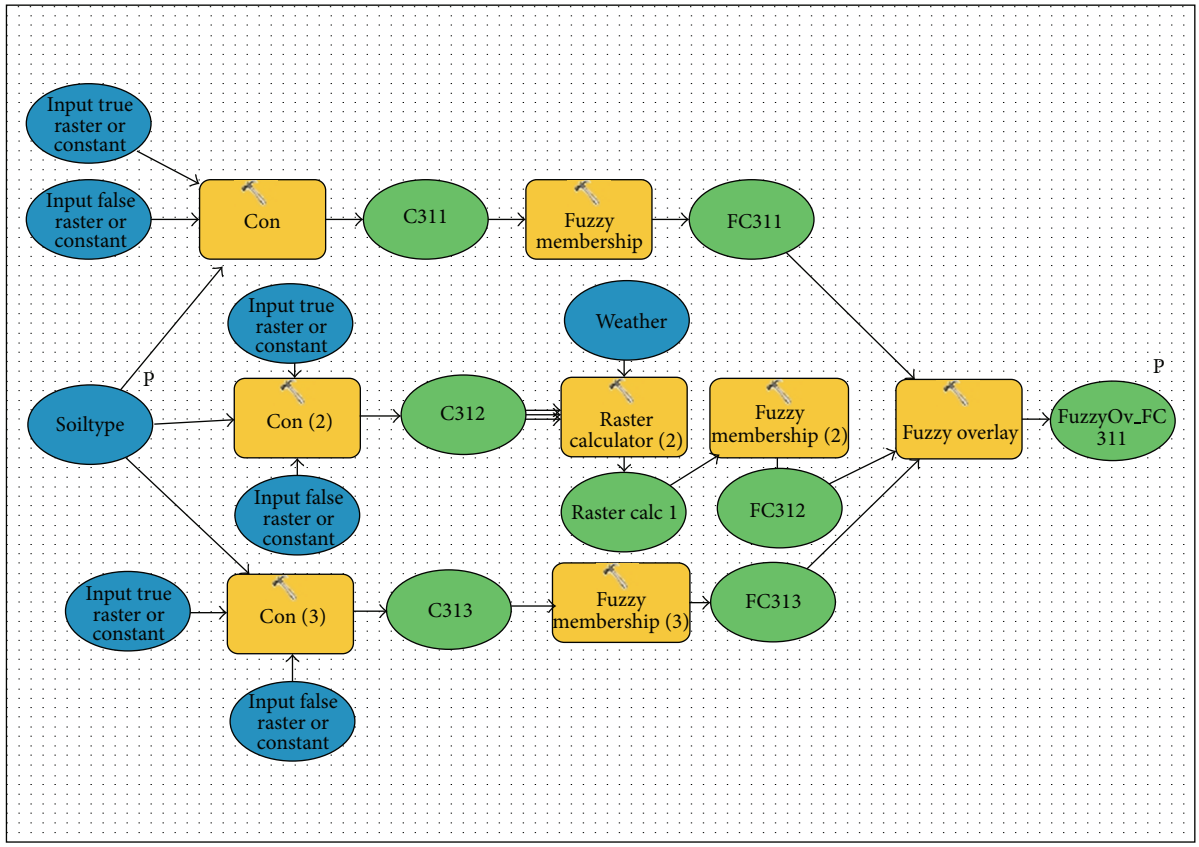

Figure 7: Demonstration of a model for calculation of deceleration coefficient created in the environment of ModelBuilder ArcGIS.

passable and this must be considered for adaptation of the calculation in such a way that the value of communication passability was more convenient for the object of communication (only communications over land are chosen) rather than for the surrounding terrain. This makes sure that for transport the route over communication is chosen and if the destination lies in a free terrain, then the route is calculated with consideration of all coefficients already on terrain. A tool of reclassification was used for calculation. In this case it is more convenient and ensures a suitable distribution of values of the attribute of transportation usage of the communication. The uncertainty in position is once again ensured by a distance raster.

The presented results are only the beginning of the solution of the complete problem. Demonstration of the fuzzy model is shown in Figure 7. Further solution will go on in two directions. They are the use of other fuzzification functions on one hand and optimization of the choice of distance, that is, vagueness level for the individual types of objects, on the other. The third problem is a determination of the final, common coefficient of deceleration and creation of a cost map. Once the third problem is successfully solved, testing of the suggested combination of parameters of fuzzification and a suggestion of methodology of decision-making with the use of fuzzy cost map will take place.

5.2. Verification of Results. The resulting cost map is only a kind of a foundation for making a decision. With the help of this map it is possible to search for an optimized route for the given vehicle type from place A to place B and judge to which extent the calculated route is suitable for the considered vehicle.

Within the solution of the project, two variants of calculation of the route for a military heavy vehicle Tatra 815 were considered [28]. One of them used a cost map derived without using fuzzy principles, and the other used these principles. With regard to the possibility to verify the calculation results, the urbanized area of the city of Brno and its closest surroundings were chosen.

With the help of these maps, routes of the stated vehicle between identical points were calculated. The results of the calculated routes are shown in the figures. With respect to the fact that for calculations reinforced communications were set as priorities, there are no significant differences in the calculated routes. Both of the calculated routes were then verified directly in real terrain. The aim of this verification was to find out to which extent the calculation itself is influenced by the use of the fuzzy method and how the use—or nonuse—of the fuzzy method will affect the given calculations. Another aim of the verification was to verify the quality of the used input data. Based on the results

of the terrain research it is possible to say that in city agglomerations where there is a sufficient net of quality reinforced communications using fuzzy principles is almost unnecessary and if sufficient quality input data are used, it is more effective to use the method of sharp boundaries for the calculation of a cost map.

Another situation, however, can happen outside urbanized areas where the net of reinforced communications is scarce and where there are a lot of forests, fields, and so forth. This fact occurred only on the edge of the researched area. That is the reason why the team for solution will focus on this type of countryside in the following steps.

Acknowledgment

The work presented in this paper was supported within the project for Development of Military Geography and Meteorology and for Support for Mathematical and Physical Research supported by the Ministry of Defence of the Czech Republic.

References

[1] ISO, *ISO 19113—Geographic Information—Quality Principles*, International Organization for Standardization, 2002, http://www.iso.org/iso/home/storecatalogue_tc/catalogue_detail.htm?csnumber=26018.

[2] ISO, *ISO 19138—Geographic Information—Data Quality Measures*, Interational Organization for Standardization, 2006, http://www.iso.org/iso/catalogue_detail.htm?csnumber=32556.

[3] V. Talhofer and A. Hofmann, "Possibilities of evaluation of digital geographic data quality and reliability," in *Proceedigs of the 24th International Cartographic Conference, the World's Geo-Spatial Solutions*, pp. 1–11, ICA/ACI, Santiago de Chilie, Chile, 2009.

[4] V. Talhofer, S. Hoskova-Mayerova, and A. Hofmann, "Improvement of digital geographic data quality," *International Journal of Production Research*, vol. 50, no. 17, pp. 4846–4859, 2012.

[5] V. Talhofer, S. Hoskova, V. Kratochvil, and A. Hofmann, "Geospatial data quality," in *International Conference on Military Technologies (ICMT '09)*, pp. 570–578, Univerzita obrany, Brno, Czech Republic, 2009.

[6] V. Kovařík, "Use of spatial modelling to select the helicopter landing sites," *Advances in Military Technology*, vol. 8, no. 2, pp. 1–10, 2013.

[7] J. D. Brown and G. B. M. Heuvelink, "The data uncertainty engine (DUE): a software tool for assessing and simulating uncertain environmental variables," *Computers & Geosciences*, vol. 33, no. 2, pp. 172–190, 2007.

[8] R. Abbaspour, R. Mahmoud, K. Delavar, and B. Reihaneh, "The issue of uncertainty propagation in spatial decision making," in *Proceedings of the 9th Scandinavian Research Conference on Geographical Information Science (ScanGIS '03)*, K. V. Tveite, Ed., pp. 57–65, Department of Surveying, Helsinki University of Technology, Espoo, Finland, 2003.

[9] M. F. Goodchild, S. Guoqing, and Y. Shiren, "Development and test of an error model for categorical data," *International Journal of Geographical Information Systems*, vol. 6, no. 2, pp. 87–104, 1992.

[10] G. Hunter, *Handling Uncertainty in Spatial Database [Ph.D. thesis]*, Department of Surveying and Land Information, University of Melbourne, Melbourne, Australia, 1993.

[11] O. Křemenová, *Fuzzy Modeling of Soil Maps*, University of Technology, Helsinky, Finland, 2004.

[12] P. Kubíček and Č. Šašinka, "Thematic uncertainty visualization usability—comparison of basic methods," *Annals of GIS*, vol. 17, no. 4, pp. 253–263, 2011.

[13] NIST, *Spatial Data Transfer Standard (FIPS 173)*, National Institute of Standards and Technology, US Department of Commerce, Washington, DC, USA, 1992.

[14] P. D'Amico, F. Di Martino, and S. Sessa, "A GIS as a decision support system for planning sustainabel mobility in a case-study," in *Multicriteria and Multiagend Decision Making with Applications to Economics and Social Sciences*, A. Ventre, A. Maturo, S. Hoskova-Mayerova, and J. Kacprzyk, Eds., Studies in Fuzziness and Soft Computing, pp. 115–128, Springer, Berlin, Germany, 2013.

[15] F. De Felice and A. Petrillo, "Decision making analysis to improve public participation in stategic energy production management," in *Multicriteria and Multiagend Decision Making with Applications to Economics and Social Sciences*, A. Ventre, A. Maturo, S. Hoskova-Mayerova, and J. Kacprzyk, Eds., Studies in Fuzziness and Soft Computing, pp. 129–142, Springer, Berlin, Germany, 2013.

[16] N. Gershon, "Visualization of an imperfect world," *IEEE Computer Graphics and Applications*, vol. 18, no. 4, pp. 43–45, 1998.

[17] A. MacEachren, "Visualizing uncertain inforamtion," *Cartographic Perspectives*, vol. 13, pp. 10–19, 1992.

[18] J. Smith, D. Retchless, C. Kinkeldey, and A. Klippel, "Beyond the surface: current issues and future irections in uncertainty visualization research," in *Proceedigs of the 26th International Cartographic Conference*, pp. 1–10, ICA, Dresden, Germany, 2013.

[19] B. Ahmad and A. Kharal, "Fuzzy sets fuzzy s-open and s-closed mappings," *Advances in Fuzzy Systems*, vol. 2009, Article ID 303042, 5 pages, 2009.

[20] F. Di Martino, V. Loia, and S. Sessa, "Fuzzy transforms method in prediction data analysis," *Fuzzy Sets and Systems*, vol. 180, no. 1, pp. 146–163, 2011.

[21] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

[22] F. Di Martino and S. Sessa, "Spatial analysis and fuzzy relation equations," *Advances in Fuzzy Systems*, vol. 2011, Article ID 429498, 14 pages, 2011.

[23] I. Cristea and S. Hoskova, "Fuzzy pseudotopological hypergroupoids," *Iranian Journal of Fuzzy Systems*, vol. 6, no. 4, pp. 11–19, 2009.

[24] ESRI, "User documentation," Copyright © 1995-2013 Esri.

[25] M. Rybansky, *Cross-Country Movement, the Imapct and Evaluation of Geographic Factors*, Akademické nakladatelství CERM, s.r.o. Brno, Brno, Czech Republic, 1st edition, 2009.

[26] M. Rybansky and M. Vala, "Relief impact on transport," in *International Conference on Military Technologies (ICMT '09)*, pp. 551–559, University of Defence, Brno, Czech Republic, 2010.

[27] MoD-GeoS, *Catalogue of the Topographic Objects DMU25*, Ministry of Defence of the Czech Republic, Geographic Service, Dobruska, Czech Republic, 7.3 edition, 2010.

[28] Tatra, *Tatra is the solution. (Tatra, a.s.)*, TATRA, 2010, http://partners.tatra.cz/exter_pr/vpnew/typovy_listprospekt.asp?kod=341&jazyk=CZ.

Hindawi Publishing Corporation
Advances in Fuzzy Systems
Volume 2013, Article ID 385974, 7 pages
http://dx.doi.org/10.1155/2013/385974

Research Article

Spatiotemporal Hotspots Analysis for Exploring the Evolution of Diseases: An Application to Oto-Laryngopharyngeal Diseases

Ferdinando Di Martino,[1] Roberta Mele,[1] Umberto E. S. Barillari,[2] Maria Rosaria Barillari,[2] Irina Perfilieva,[3] and Sabrina Senatore[4]

[1] *Universitá degli Studi di Napoli Federico II, Dipartimento di Architettura, Via Toledo 402, 80134 Napoli, Italy*
[2] *Seconda Università degli Studi di Napoli, Dipartimento di Psichiatria, Neuropsichiatria Infantile, Audiofoniatria e Dermatovenereologia, L.go Madonna delle Grazie, 80138 Napoli, Italy*
[3] *Centre of Excellence IT4 Innovations, Institute for Research and Applications of Fuzzy Modelling, University of Ostrava, 30. dubna 22, 70103 Ostrava, Czech Republic*
[4] *Universitá degli Studi di Salerno, Dipartimento di Informatica, Via Ponte don Melillo, 80084 Fisciano, Salerno, Italy*

Correspondence should be addressed to Ferdinando Di Martino; fdimarti@unina.it

Received 1 May 2013; Revised 29 July 2013; Accepted 29 July 2013

Academic Editor: Salvatore Sessa

This paper presents a spatiotemporal analysis of hotspot areas based on the Extended Fuzzy C-Means method implemented in a geographic information system. This method has been adapted for detecting spatial areas with high concentrations of events and tested to study their temporal evolution. The data consist of georeferenced patterns corresponding to the residence of patients in the district of Naples (Italy) to whom a surgical intervention to the oto-laryngopharyngeal apparatus was carried out between the years 2008 and 2012.

1. Introduction

In a GIS, the impact of phenomena in a specific area due to the proximity of the event (e.g., the study of the impact area of an earthquake, or the area constraint around a river basin) is performed using buffer area geoprocessing functions. Given a geospatial event topologically represented as a georeferenced punctual, linear, or areal element, an atomic buffer area is constituted by circular areas centered on the element. For example, if the event is the epicenter of an earthquake, georeferenced by a point, a set of buffer areas is formed by concentric circular areas around that point; the radius of each circular buffer area is defined a priori.

When it is not possible to define statically an area of impact and we need to determine what is the area affected by the presence of a consistent set of events, we are faced with the problem of detecting this area as a cluster on which the georeferenced events are thickened as well. These clusters are georeferenced, represented as polygons on the map, and called *hotspot areas*.

The study of hotspot areas is vital in many disciplines such as crime analysis [1–3], which studies the spread on the territory of criminal events, fire analysis [4], which analyzes the phenomenon of spread of fires on forested areas, and disease analysis [5–7], which studies the localization of focuses of diseases and their temporal evolution. The clustering methods mainly used for detecting hotspot areas are the algorithms based on density (see [8, 9]); they can detect the exact geometry of the hotspots, but are highly expensive in terms of computational complexity, and in the great majority of cases, it is not necessary to determine exactly the shape of the clusters. The clustering algorithm more used for its linear computational complexity is the Fuzzy C-Means algorithm (FCM) [10], a partitive fuzzy clustering method that uses the Euclidean distance to determine prototypes cluster as points.

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset R^n$ be a dataset composed of N pattern $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T$, where x_{kj} is the kth component (feature) of the pattern $\mathbf{x}_j$. The FCM algorithm minimizes the following objective function:

$$J(\mathbf{X}, \mathbf{U}, \mathbf{V}) = \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^m d_{ij}^2, \tag{1}$$

where C is the number of clusters, fixed a priori, u_{ij} is the membership degree of the pattern $\mathbf{x}_j$ to the ith cluster ($i = 1, \ldots, C$), $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_C\} \subset R^n$ is the set of points given by the centers of the C clusters (prototypes), m is the fuzzifier parameter and, d_{ij} is the distance between the center $\mathbf{v}_i = (v_{1i}, v_{2i}, \ldots, v_{ni})^T$ of the ith cluster and the jth vector $\mathbf{x}_j$, calculated as the Euclidean norm:

$$d_{ij} = \left\| \left(\mathbf{x}_j - \mathbf{v}_i \right) \right\| = \sqrt{\sum_{k=1}^{n} \left(\mathbf{x}_{kj} - \mathbf{v}_{ki} \right)^2}. \tag{2}$$

Using the Lagrange multipliers method for minimizing the objective function (1), we obtain the following solution for the center of each cluster prototype:

$$\mathbf{v}_i = \frac{\sum_{j=1}^{N} u_{ij}^m \mathbf{x}_j}{\sum_{j=1}^{N} u_{ij}^m}, \tag{3}$$

where $i = 1, \ldots, C$ and for membership degrees u_{ij},

$$u_{ij} = \frac{1}{\left(\sum_{h=1}^{c} \left(d_{ij}^2 / d_{hj}^2 \right) \right)^{2/(m-1)}} \tag{4}$$

subjected to the constraints:

$$\begin{aligned} &\sum_{i=1}^{C} u_{ij} = 1, \quad \forall j \in \{1, \ldots, N\}, \\ &0 < \sum_{j=1}^{N} u_{ij} < N, \quad \forall i \in \{1, \ldots, C\}. \end{aligned} \tag{5}$$

Initially, the u_{ij}'s and the $\mathbf{v}_i$ are assigned randomly and updated in each iteration. If $U^{(l)} = (u_{ij}^{(l)})$ is the matrix $\mathbf{U}$ calculated at the l-th step, the iterative process stops when

$$\left\| \mathbf{U}^{(l)} - \mathbf{U}^{(l-1)} \right\| = \max_{i,j} \left| u_{ij}^{(l)} - u_{ij}^{(l-1)} \right| < \varepsilon, \tag{6}$$

where $\varepsilon > 0$ is a prefixed parameter.

This algorithm has a linear computational complexity; however, it is sensitive to the presence of noise and outliers; furthermore, the number of cluster C is fixed a priori and needs to use a validity index for determining an optimal value for the parameter C.

In order to overcome these shortcomings, in [11, 12], the EFCM algorithm is proposed, where the cluster prototypes are hyperspheres in the case of the Euclidean metric. Like FCM, the EFCM algorithm is characterized by a linear computational complexity; furthermore, it is robust with respect to the presence of noise and outliers, and the final number of clusters is determined during the iterative process.

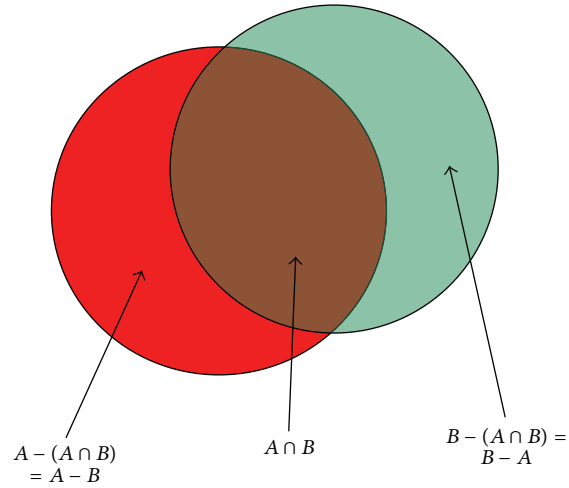

Figure 1: Intersections of two hotspots detected for events that happened in two consecutive periods.

In [13, 14], the authors propose the use of the EFCM algorithm for detecting hotspot areas. The final hotspots are identified as the detected cluster prototypes and shown on the map as circular areas. In [4], the authors analyze the spatio-temporal evolution of the hotspots in the fire analysis. The pattern event dataset is partitioned according to the time of the event's detection; so each subset is corresponding to a specific time interval. The authors compare the hotspots obtained in two consecutive years by studying their intersections on the map. In this way, it is possible to follow the evolution of a particular phenomenon.

The cluster prototypes detected from EFCM method are circular areas on the map that can approximate a hotspot area. Figure 1 shows an example of two circular hotspots, obtained as clusters.

Figure 1 shows three different regions.

(i) An area in which the hotspot A is not intersected by the hotspot B (corresponding to $A - (A \cap B) = A - B$): this region can be considered as a geographical area in which prematurely detected event disappears successively.

(ii) The region of intersection of the two hotspots $A \cap B$: this region can be considered a geographical area in which the event continues to persist.

(iii) An area in which the hotspot B is not intersected by the hotspot A (corresponding to $B - (A \cap B) = B - A$): this region can be considered as a geographical area in which the prematurely undetected event propagates successively.

We can study the spatio-temporal evolution of the hotspots by analyzing the interactions between the corresponding circular cluster prototypes obtained for consecutive periods,

and detecting the presence of new hotspots in regions previously not covered by hotspots and the absebce of hotspots in regions previously spatially included in hotspot areas.

In this research, we present a method for studying the spatio-temporal evolution of hotspots areas in disease analysis; we apply the EFCM algorithm for comparing, in consecutive years, event datasets corresponding to otolaryngopharyngeal diseases diagnosis detected in the district of Naples (I). Each event corresponds to the residence of the patient who contracted the disease.

We study the spatio-temporal evolution of the hotspots analyzing the intersections of hotspots corresponding to two consecutive years, the displacement of the centroids, the increase or reduction of the hotspots areas, and the emergence of new hotspots.

In Section 2, we give an overview of the EFCM algorithm. In Section 3, we present our method for studying the spatio-temporal evolution of hotspots in disease analysis. In Section 4, we present the results of the spatio-temporal evolution of hotspots for the otolaryngologist-laryngopharyngeal diseases diagnosis events detected in the district of Naples (I). Our conclusions are in Section 5.

2. The EFCM Algorithm

In the EFCM algorithm, we consider clustering prototypes given by hyperspheres in the n-dimensional feature's space. The ith hypersphere is characterized by a centroid $\mathbf{v}_i = (\mathbf{v}_{1i}, \ldots, \mathbf{v}_{ni})$ and a radius r_i.

Indeed, if r_i is the radius of V_i, we say that x_j belongs to V_i if $d_{ij} \leq r_i$.

The radius r_i is obtained considering the covariance matrix P_i associated with the ith cluster, defined as

$$\mathbf{P}_i = \frac{\sum_{j=1}^{N} u_{ij}^m \left(\mathbf{x}_j - \mathbf{v}_i\right)\left(\mathbf{x}_j - \mathbf{v}_i\right)^T}{\sum_{j=1}^{N} u_{ij}^m} \tag{7}$$

whose determinant gives the volume of the ith cluster. Since $\mathbf{P}_i$ is symmetric and positive, it can be decomposed in the following form:

$$\mathbf{P_i} = \mathbf{Q}_i \mathbf{\Lambda}_i \mathbf{Q}_i^T, \tag{8}$$

where Q_i is an orthonormal matrix and $\Delta_i = (\lambda_{ij})$ is a diagonal matrix. The radius r_i is given by the following formula (see [12]):

$$r_i = \frac{1}{n}\sqrt{\prod_{k=1}^{n} \lambda_{ik}^{1/n}} = \sqrt{\det\left(P_i\right)^{1/n}}. \tag{9}$$

The objective function to be minimized is the following:

$$J\left(X, U, V\right) = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}^m \left(d_{ij}^2 - r_i^2\right), \tag{10}$$

where the membership degrees u_{ij} are updated as

$$\begin{aligned} u_{ij} &= 1 \times \left(\sum_{h=1}^{C} \left(\left(d_{ij}^2 \max\left(0, 1 - r_i^2/d_{ij}^2\right)\right) \right.\right. \\ &\qquad \left.\left. / \left(d_{hj}^2 \max\left(0, 1 - r_h^2/d_{hj}^2\right)\right)\right)^{1/(m-1)} \right)^{-1} \\ &= \frac{1}{\sum_{h=1}^{C} \left(d_{ij}^2 w_{ij} / d_{hj}^2 w_{hj} \right)^{1/(m-1)}}. \end{aligned} \tag{11}$$

We set $d_{kj}'^2 = d_{kj}^2 w_{kj} = \max(0, d_{kj}^2 - r_k^2)$ and define the number $\varphi_j = \text{card}\{k \in \{1, \ldots, C\} : d_{kj}' = 0\}$ for any $j = 1, \ldots, N$; thus, we obtain

$$\begin{aligned} u_{ij} &= \frac{1}{\sum_{h=1}^{C} \left(d_{ij}' / d_{hj}' \right)^{2/(m-1)}} \quad \text{if } \varphi_j = 0, \\ u_{ij} &= \begin{cases} 0 & \text{if } d_{ij}' > 0, \\ & \text{if } \varphi_j > 0. \\ \dfrac{1}{\varphi_j} & \text{if } d_{ij}' = 0, \end{cases} \end{aligned} \tag{12}$$

However, the usage of (12) produces the negative effect of diminishing the objective function (10) when a meaningful number of features are placed in a cluster and this fact can prevent the separation of the clusters. Then a solution to this problem consists in the assumption of a small starting value of r_i and then it is increased gradually with the factor $\beta^{(l)}/C^{(l)}$, where $C^{(l)}$ is the number of clusters at the lth iteration and $\beta^{(l)}$ is defined recursively as $\beta^{(0)} = 1$, $\beta^{(l)} = \min(C^{(l-1)}, \beta^{(l-1)})$, by setting

$$I_{ik} = \frac{\sum_{j=1}^{N} \min\left(u_{ij}, u_{ki}\right)}{\sum_{j=1}^{N} u_{ij}} \tag{13}$$

and the symmetric matrix $S = (S_{ik})$, where $S_{ik} = \max\{I_{ik}, I_{ki}\}$ is defined as well. If $S^{(l)}$ is the matrix S at the lth iteration ($l > 1$) and the threshold $\alpha^{(l)} = 1/(C^{(l)} - 1)$ is introduced as limit, then two indexes i^* and k^* are determined such that $S_{i^*k^*}^{(l)} \geq \alpha^{(l)}$ and thus i^* and k^* are merged by setting

$$\begin{aligned} u_{i^*j}^{(l)} &= u_{i^*j}^{(l)} + u_{k^*j}^{(l)}, \quad \forall j \in \{1, \ldots, N\}, \\ C^{(l)} &= C^{(l-1)} + 1. \end{aligned} \tag{14}$$

The k^*th row can be removed from the matrix $U^{(l)}$. In other words, the EFCM algorithm can be summarized in the following steps.

(1) The user assigns the initial number of clusters $C^{(0)}$, $m > 1$ (usually $m = 2$), $\varepsilon > 0$, the initial value $S_{ik}^{(0)} = 0$, and $\beta^{(0)} = 1$.

(2) The membership degrees $u_{ij}^{(0)}$ ($j = 1, \ldots, N$ and $i = 1, \ldots, C(0)$) are assigned randomly.

Figure 2: Example of events georeferenced on a road network.

(3) The centers of the clusters v_i are calculated by using (3).

(4) The radii of the clusters are calculated by using (9).

(5) u_{ij} is calculated by using (12).

(6) The indexes i^* and k^* are determined in such a way that $S^{(l)}_{i^*k^*}$ assumes the possible greatest value at the lth iteration.

(7) If $|S^{(l)}_{i^*k^*} - S^{(l-1)}_{i^*k^*}| < \varepsilon$ and $S^{(l)}_{i^*k^*} > \alpha(l) = 1/(C^{(l-1)} - 1)$, then the i^*th and k^*th clusters are merged via (14) and the k^*th row is deleted from $U^{(l)}$.

(8) If (6) is satisfied, then the process stops; otherwise, go to the step (3) for the $(l + 1)$th iteration.

3. Hotspots Detection and Evolution in Disease Analysis

Each pattern is given by the event corresponding to the residence of the patient to whom a specific disease has been detected. The two features of the pattern are the geographic coordinates of the residence.

The first step of our process is a geocoding activity necessary for obtaining the event dataset starting by the street address of the patients.

To ensure an accurate matching for the geopositioning of the event, we need the topologically correct road network and the corresponding complete toponymic data.

The starting data include the name of the street and the house number of the patient's residence. After the matching process, each data is converted in an event point georeferenced on the map.

In Figure 2, the road network of the district of Naples is shown; the name of the street is labeled on the map; the events are georeferenced as points on the map.

Figure 3 shows the data corresponding to an event selected on the map.

After geo-referencing each event, the event dataset can be split, partitioning them by time interval. For example, the event in Figure 3 can be split by the field "Year."

For each subset of events, we apply the EFCM algorithm to detect the final cluster prototypes.

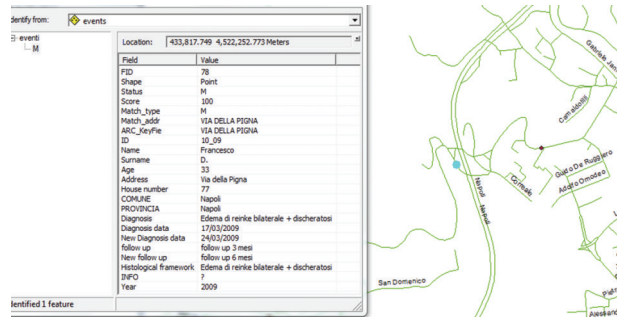

Figure 3: Data associated to an event on the map.

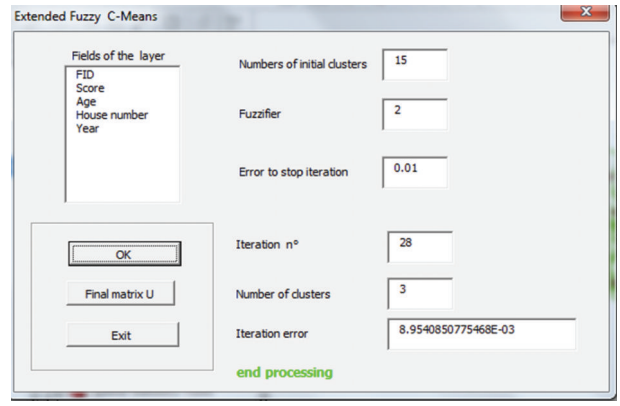

Figure 4: A form created in the GIS Tool ESRI ArcGIS for managing the EFCM process.

In this research, we point out the analysis of the temporal evolution and spread of oto-laryngo-pharyngeal diseases detected within the district of Naples. The datasets, divided by time sequences corresponding to periods of one year, are made up of patterns for different events georeferenced corresponding to ailments encountered in patients for which an intervention and the subsequent histological examination were pointed out as well. The event refers to the geopositioning of the location of the patient.

The data have been further divided by the type of the disease for analyzing the distribution and evolution of each specific disease on the area of the study.

The EFCM algorithm has been encapsulated in the GIS platform ESRI ArcGIS. Figure 4 shows the mask created for setting the parameters and running the EFCM algorithm.

We can set other numerical fields for adding other features to the geographical coordinates.

Initially, we set the initial number of clusters, the fuzzifier m, and the error threshold for stopping the iterations. After running EFCM, the number of iterations, the final number of clusters, and the error calculated at the last iteration are reported. The resultant clusters are shown as circular areas on the map and can be saved in a new geographic layer.

The final process concerns the comparative analysis of the hotspots obtained by the clusters corresponding to each subsets of events.

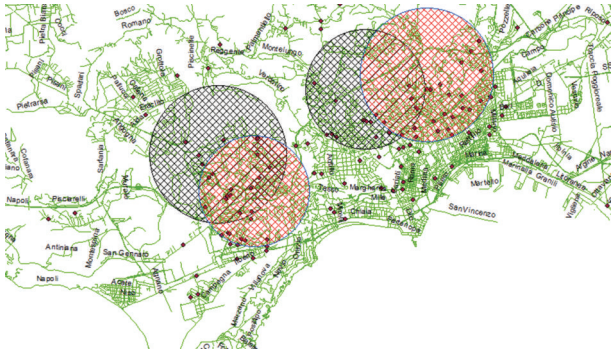

Figure 5: Analysis of the spatio-temporal evolution of hotspots detected in two consecutive periods.

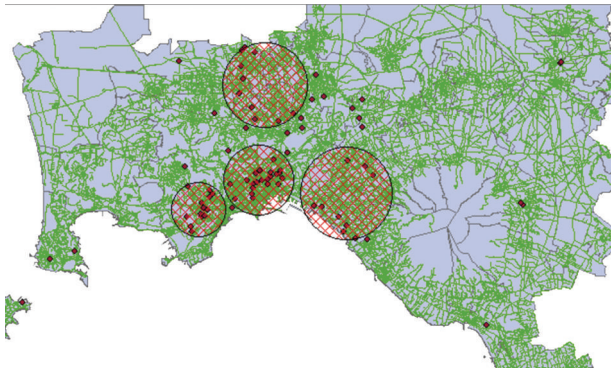

Figure 6: Edema of bilateral Reinke disease—year 2011: display of the hotspots on the map.

Figure 5 shows an example of display on the map of hotspots obtained as final clusters for two consecutive subset of events.

In order to assess the expansion and the displacement of a hotspot, we measure the radius of the hotspot and the distance between the centroids of two intersecting hotspots.

In the next section, we present the results obtained by applying this method for the data corresponding to surgical interventions to the oto-laryngo-pharyngeal apparatus in patients residents in the district of Naples between the years 2008 and 2012.

We divide the dataset per year and analyze various types of diseases.

Among the types of the most frequent diseases, the following were analyzed:

(i) carcinoma,

(ii) edema of bilateral Reinke,

(iii) hypertrophy of the inferior turbinate,

(iv) nasal polyposis,

(v) bilateral vocal fold prolapse.

In the next section, we show the most significant results obtained by applying this method to the each partitioned dataset of events.

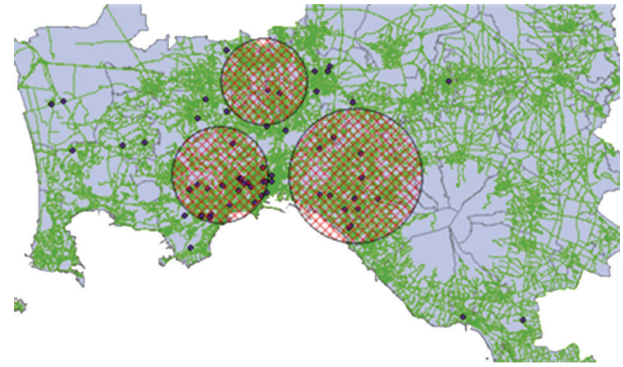

Figure 7: Edema of bilateral Reinke disease—year 2012: display of the hotspots on the map.

Table 1: Endema of bilateral Reinke disease—final number of clusters and final error for year.

Year	Initial number of clusters	Final number of clusters	$\lvert U^{(l)} - U^{(l-1)} \rvert$	ε
2008	15	4	0.48×10^{-2}	1×10^{-2}
2009	15	4	0.55×10^{-2}	1×10^{-2}
2010	15	4	0.71×10^{-2}	1×10^{-2}
2011	15	4	0.67×10^{-2}	1×10^{-2}
2012	15	3	0.53×10^{-2}	1×10^{-2}

4. Test Results

We present the results obtained on the event dataset described above in the period between the years 2008 and 2012.

We consider first the subset of data corresponding to the edema of bilateral Reinke disease.

We fix the fuzzifier parameter to 0.1, the initial number of clusters to 15, and the final iteration error to 1×10^{-2}.

Table 1 shows the results obtained for each year.

We present the details relating to the comparison of the hotspots obtained by considering the event data for the years 2011 and 2012.

Figures 6 and 7 show, respectively, the hotspots obtained by using the pattern subset of events that occurred in the years 2011 and 2012.

Figure 8 shows the overlap of the hotspots obtained for the two years: in red, the hotspots corresponding to the year 2011; in blue, the ones corresponding to the year 2012.

Table 2 shows in the first two columns the labels of the hotspots in 2011 and 2012, in third (resp., fourth) column the radius obtained in 2011 (resp., 2012), and the distance between the centroids is given in the fifth column.

The results show that only hotspot 3 obtained for the year 2011 remains almost unchanged in the year 2012. Instead, hotspots 1 and 2 seem to merge into a single larger hotspot (the hotspot 1 obtained for the year 2012), and hotspot 4, that shifts about 1 km, is expanded; the radius of this hotspot in 2012 is about 6.5 km (hotspot 3 obtained for the year 2012 in Figure 8).

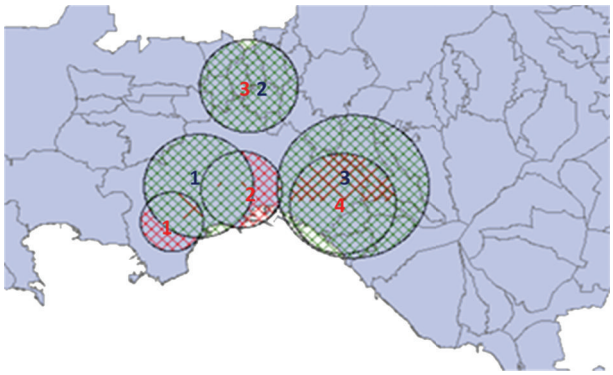

Figure 8: Edema of bilateral Reinke disease—years 2011 and 2012: display of the two hotspots' series.

Table 2: Endema of bilateral Reinke disease—comparison results of the hotspots obtained for the years 2011 and 2012.

2011 hotspot	Intersecting 2012 hotspot	Radius 2011 hotspot (km)	Radius 2012 hotspot (km)	Centroid's distance (km)
1	1	1.724	3.848	1.759
2	1	1.943	3.848	1.507
3	2	3.434	3.453	0.074
4	3	3.591	6.519	1.115

Table 3: Nasal polyposis—comparison results of the hotspots obtained for the years 2011 and 2012.

2011 hotspot	Intersecting 2012 hotspot	Radius 2011 hotspot (km)	Radius 2012 hotspot (km)	Centroid's distance (km)
1	1	3.087	4.951	2.656
2	2	4.915	7.103	1.052

Now we show the results obtained for the disease nasal polyposis.

Figure 9 shows the overlap of the hotspots obtained for the two years, 2011 and 2012.

In Table 3, the comparison's results are reported.

The results in Figure 9 show that in 2011 and 2012 there are two hotspots: the one covering an area of the city of Naples and the other covering many Vesuvian towns. The two hotspots, which in 2011 covered a circular area with a radii of about 3 and 5 km, respectively, in 2012 cover a circular area with radii of about 5 and 7 km, respectively.

The histogram in Figure 10 shows the trend of the radii of the two hotspots in the course of time.

It is relevant the spread in recent years of the hotspot that surrounds the Vesuvian towns (the radius of this hotspot, from about 2 km in the year 2008, is about 7 km in the year 2012).

Another significant trend concerns the hotspots obtained for the carcinoma disease.

Also, in this case, the two main hotsposts cover the city of Naples and many Vesuvian towns. In in this case, we have a very high spread of the hotspot covering the city of Naples

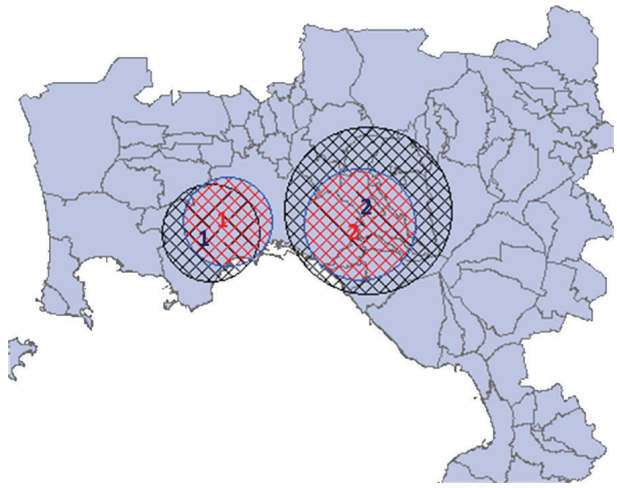

Figure 9: Nasal polyposis disease—years 2011 and 2012: display of the two hotspots' series.

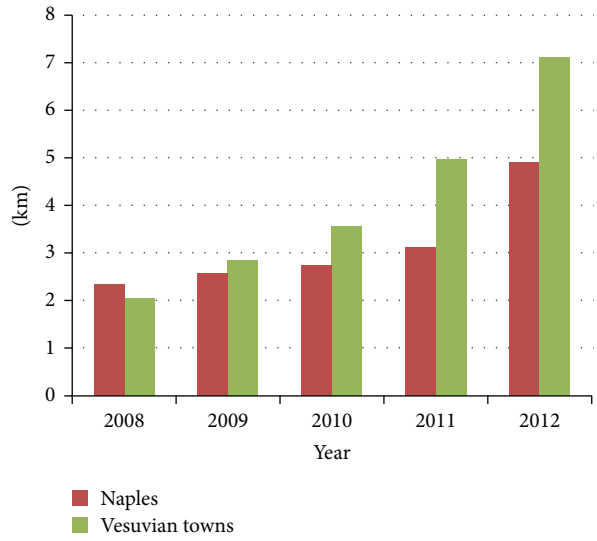

Figure 10: Nasal polyposis disease—histogram showing the variation of the radius of the two hotspots over time.

(cfr. Figure 11); in recent years, the radius of this hotspot is increased up to 9.5 km.

5. Conclusions

The hyperspheres obtained as clusters (circles in case of two dimensions) by using EFCM can represent hotspots in hotspot analysis; this method has a linear computational complexity and is robust to noises and outliers. In hotspots analysis, the patterns are bidimensional and the features are formed by geographic coordinates; the cluster prototypes are circles that can represent a good approximation of hotspot areas and can be displayed as circular areas on the map.

In this paper, we present a new method that uses the EFCM algorithm for studying the spatio-temporal evolution of hotspots in disease analysis.

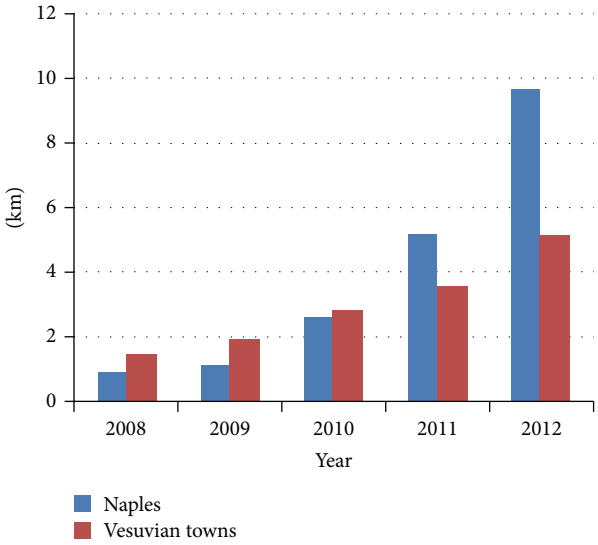

Figure 11: Carcinoma disease—histogram showing the variation of the radius of the two hotspots over time.

We consider the residence's information of patients in the district of Naples (Italy) to whom a surgical intervention to the oto-laryngo-pharyngeal apparatus was carried out between the years 2008 and 2012. A geocoding process is used for geo-referencing the data; then, the georeferenced dataset is partitioned per year and type of disease; we compare the hotspots obtained for each pair of consecutive years and analyze the trend of each hotspot over time measuring the variation of the radius and the distance between intersecting cluster centroids concerning two consecutive years.

The results show a consistent spread in the last years of the nasal polyposis disease hotspot covering some Vesuvian towns and of the carcinoma disease hotspot covering the city of Naples.

References

[1] S. P. Chainey, S. Reid, and N. Stuart, "When is a hotspot a hotspot? A procedure for creating statistically robust hotspot geo-graphic maps of crime," in *Innovations in GIS 9: Socioeconomic Applications of Geographic Information Science*, D. Kidner, G. Higgs, and S. White, Eds., Taylor and Francis, London, UK, 2002.

[2] K. Harries, *Geographic Mapping Crime: Principle and Practice*, National Institute of Justice, Washington, DC, USA, 1999.

[3] A. T. Murray, I. McGuffog, J. S. Western, and P. Mullins, "Exploratory spatial data analysis techniques for examining urban crime," *British Journal of Criminology*, vol. 41, no. 2, pp. 309–329, 2001.

[4] F. Di Martino and S. Sessa, "The extended fuzzy c-means algorithm for hotspots in spatio-temporal GIS," *Expert Systems with Applications*, vol. 38, no. 9, pp. 11829–11836, 2011.

[5] R. M. Mullner, K. Chung, K. G. Croke, and E. K. Mensah, "Introduction: geographic information systems in public health and medicine," *Journal of Medical Systems*, vol. 28, no. 3, pp. 215–221, 2004.

[6] K. Polat, "Application of attribute weighting method based on clustering centers to discrimination of linearly non-separable medical datasets," *Journal of Medical Systems*, vol. 36, no. 4, pp. 2657–2673, 2012.

[7] C. K. Wei, S. Su, and M. C. Yang, "Application of data mining on the develoment of a disease distribution map of screened community residents of Taipei County in Taiwan," *Journal of Medical Systems*, vol. 36, no. 3, pp. 2021–2027, 2012.

[8] I. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 773–780, 1989.

[9] R. Krishnapuram and J. Kim, "Clustering algorithms based on volume criteria," *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 2, pp. 228–236, 2000.

[10] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, NY, USA, 1981.

[11] U. Kaymak, R. Babuska, M. Setnes, H. B. Verbruggen, and H. M. van Nauta Lemke, "Methods for simplification of fuzzy models," in *Intelligent Hybrid Systems*, D. Ruan, Ed., pp. 91–108, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

[12] U. Kaymak and M. Setnes, "Fuzzy clustering with volume prototypes and adaptive cluster merging," *IEEE Transactions on Fuzzy Systems*, vol. 10, no. 6, pp. 705–712, 2002.

[13] F. Di Martino, V. Loia, and S. Sessa, "Extended fuzzy c-means clustering algorithm for hotspot events in spatial analysis," *International Journal of Hybrid Intelligent Systems*, vol. 4, pp. 1–14, 2007.

[14] F. Di Martino and S. Sessa, "Implementation of the extended fuzzy c-means algorithm in geographic information systems," *Journal of Uncertain Systems*, vol. 3, no. 4, pp. 298–306, 2009.