Fuzzy models in regional statistics

Sunanta, Owat and Viertl, Reinhard

Technische Universität Wien

2016

Fuzzy models in regional statistics

Owat Sunanta
Technische Universität Wien,
Vienna, Austria
E-mail:
owat.sunanta@tuwien.ac.at

Reinhard Viertl
Technische Universität Wien,
Vienna, Austria
E-mail:
r.viertl@tuwien.ac.at

Many regional data are not provided as precise numbers, but they are frequently non-precise (fuzzy). In order to provide realistic statistical information, the imprecision must be described quantitatively. This is possible using special fuzzy subsets of the set of real numbers $\mathbb{R}$, called fuzzy numbers, together with their characterising functions. In this study, the uncertainty of measured data is highlighted through an example of environmental data from a regional study. The generalised statistical methods, through the characterising function and the δ-cut, that are suitable for the situations of fuzzy uni- and multivariate data are described. In addition, useful generalised descriptive statistics and predictive models frequently applicable for analysis of fuzzy data in regional studies as well as the concept of fuzzy data in databases are presented.

Keywords:
fuzzy data in regional studies,
characterising function,
statistics with fuzzy data,
fuzzy data in databases

Introduction

The measurement of continuous variables is often clouded with uncertainty, and many data are not exact numbers but more or less fuzzy. This type of uncertainty is different from errors. However, the count data are, if considered in a larger scope, often associated with various types of uncertainty. The inaccuracy of data is not usually assumed in standard statistics. However, these problems should be approached with caution. Inaccurate data are quite general in many environmental analyses, and occur often in regional studies. In these cases, the inaccurate data are often presented with considerable uncertainties. Nevertheless, such data are essential for decisions, despite their uncertainties. Lee (1995) proposed some useful concepts of fuzzy spatial statistics. The work by Burrough (2001) emphasises that the fuzzy set theory is a useful tool for spatial analysis. There have also been efforts to apply fuzzy models in the field, for example, for the assessment of urban air quality (Guleda–Ibrahim–Halil–2004) and the estimation of underground economy (Ene–Hurduc 2010).

Regional Statistics, Vol 6, No 1. 2016: 104–118; DOI: 10.15196/RS06106

The description of fuzzy data and their statistical analysis also form an active field of research. The most suitable mathematical model to describe the fuzziness is fuzzy numbers and their characterising functions (Viertl 2015). In this contribution, the generalised statistical methods to handle fuzzy data, usually found in regional studies, are described. In the next section definition and examples of fuzzy data are provided. Characterisation of fuzzy data through special membership functions of fuzzy numbers, i.e. so-called characterising functions, is described in the third section. Some useful descriptive statistics for fuzzy data are explained in the fourth section. In the fifth section models for prediction based on fuzzy information are described. In the sixth section, the use of fuzzy data in databases is introduced. The contribution is concluded with final remarks in the final section.

Fuzzy Data in Regional Studies

In regional studies, measurements, often statistical, are necessary for further analyses. The concept of measurement has been developed in conjunction with the concepts of numbers and units of measurement. In statistics, data as a result of measurements are typically categorised at different levels, i.e. nominal, ordinal, interval and ratio data. Knowing the level of the measurement helps in applying appropriate methods and/or models in interpreting and analysing data of different levels accordingly.

Examples of one-dimensional fuzzy data are height of a tree, water levels in lakes or rivers and concentrations of toxic substances in the air. On the other hand, many measurements under consideration are generally in the form of multivariate data (Wichern–Johnson 2007); that is, the corresponding idealised results are real vectors $\mathbf{x} = (x_1, \ldots, x_k) \in \mathbb{R}^k$. For example, data on several variables are used altogether in identifying factors that are responsible for a nation's growth index. These data are frequently represented in the form of time series, which requires specific methods for further data analyses, such as those introduced in the fifth section.

Real observations x of continuous stochastic quantities X are not precise numbers or vectors, whereas the measurement results are more or less non-precise, or fuzzy. The fuzziness of individual measurement results is described by the so-called fuzzy numbers, while the variability is described by stochastic models. As a result, the analysis of repeated measurements is possible by using suitably generalised statistical methods (Viertl 2015).

As an illustration, the measurement results of substances in the air are generally reported by different regions as part of the environmental data in regional studies. The Austrian Ambient Air Quality Protection Act has established air quality limit values for sulphur dioxide (SO_2), nitrogen dioxide (NO_2), nitric oxide (NO), lead, benzene, carbon monoxide (CO) and particulate matter (PM), as well as target values for ozone (Austria's Federal Environment Agency 2002). Limit values for NO_2 are often exceeded in agglomerations, predominantly at traffic-related areas. Table 1 shows the measurement of NO_2 emission in the air at different measuring stations in South Tyrol, Italy (Landesinstitut für Statistik Bozen, Südtirol 2015). Table 2 lists the levels of severity in 'linguistic' terms as an interpretation of the numerical measurement (adapted from (Amt der Tiroler Landesregierung 2016)).

Table 1

The amount* of NO2 emission in the air at different measuring stations

Stations	2009	2010	2011	2012	2013	2014
Bozen 6	...	...	...	28	32	30
Bozen 4	46	44	46	43	43	41
Bozen 5	41	40	42	39	40	37
Leifers	27	28	28	27	27	25
Meran	34	34	34	31	33	31
Latsch	18	18	19	17	18	17
Bruneck	21	22	21	20	20	19
Sterzing	32	34	34	30	31	30
Brixen	29	29	30	27	27	30
Feldthurns (A22)	67	67	65	60	60	58
Auer (A22)	49	45	47	45	45	42
Kurtinig a.d.W.	...	33	40	32	32	30
Ritten	4	3	5	3	3	3

* Annual average in $\mu g/m^3$ of the daily averages from the concentration data collected for over a year.
Source: Landesinstitut für Statistik (ASTAT), Bozen, Südtirol (2015).

Table 2

Evaluation of the level of NO2 emission in the air

Level	NO_2[a)]
Very small polluted	< 50
Small polluted	≥ 50
Polluted	≥ 80

a) Measured in $\mu g/m^3$
Source: Abteilung Raumordnung-Statistik, Amt der Tiroler Landesregierung (2002).

Measurement of a continuous variable, the amount of NO_2 in this case, is often a source of uncertainty. On the other hand, allocation of the quantitative measures into different classes is frequently necessary. The allocation itself, as well as the interpretation between classes, is defined by qualitative data (linguistic terms), and the corresponding quantitative measurement values, as in Table 2, are subjective and uncertain (fuzzy). Problems often arise when the quantitative measures lie somewhere close to and/or on the border lines between neighbouring classes, for example, the amount of NO_2 measured in Auer (A22) from Table 1. Categorising such border-line amounts as very small or small polluted can be subjective and uncertain. Subjectively categorising such measures may subsequently trigger unnecessary corrective actions, which, in turn, cost time and money. Thus, to further analyse such data appropriately, a generalised model to quantify such uncertainty is necessary.

The best (to-date) mathematical description (see also (Klir–Yuan 1995)) of such data (observations) is by means of fuzzy numbers $x_1^*, \dots, x_n^*$ with corresponding characterising functions $\xi_1(\cdot), \dots, \xi_n(\cdot)$, described in the next section or by a fuzzy vector $\mathbf{x}^*$ with corresponding vector-characterising function $\zeta(\cdot, \dots, \cdot)$ for multivariate fuzzy data.

Characterisation of fuzzy data

In order to describe observations or measurements of continuous quantities, the definition of general fuzzy numbers is useful.

Definition 1: A general fuzzy number x^* is defined by its characterising function $\xi(\cdot)$, which is a real function of one real variable and has the following properties:

(1) $\xi: \mathbb{R} \to [0,1]$

(2) The support of $\xi(\cdot)$, denoted by $supp[\xi(\cdot)]$ and defined by
$supp[\xi(\cdot)] \coloneqq \{x \in \mathbb{R}: \xi(x) > 0\}$,
is a bounded subset of $\mathbb{R}$.

(3) For all $\delta \in (0,1]$, the δ-cut $C_\delta[\xi(\cdot)]$, defined by
$C_\delta[\xi(\cdot)] \coloneqq \{x \in \mathbb{R}: \xi(x) \geq \delta\} = \bigcup_{j=1}^{k_\delta} [a_{\delta,j}, b_{\delta,j}]$,
is non-empty and finite union of compact intervals.

Along with general fuzzy numbers, a related critical question is how to obtain the characterising function of a measurement result. First, a function has to be defined and then the special membership functions of fuzzy numbers, which are characterising functions, describing measurement results, can be obtained (Kovářová–Viertl 2015).

Observations or measurements of continuous quantities obtained from the measuring equipment are norms in regional studies. In case of analogue measuring equipment, the measurement result can be read from a pointer position on a scale and further recorded by a photograph. Such photographs display the position of the reading pointer in the form of colour intensity $g(\cdot)$ along a measurable scale. The characterising function $\xi(\cdot)$ can then be obtained in the following way:

Taking the value c of the basic colour intensity, a function $h(\cdot)$ is defined as

$$h(x) \coloneqq g(x) - c \qquad \forall x \in \mathbb{R}.$$

Based on the function $h(\cdot)$, the characterising function $\xi(\cdot)$ of the fuzzy number describing the measurement is obtained in the following way:

$$\xi(x) \coloneqq \frac{|h'(x)|}{max\{|h'(x)|: x \in \mathbb{R}\}} \qquad \forall x \in \mathbb{R},$$

where $h'(\cdot)$ is the derivative of function $h(\cdot)$.

For measurements of vector quantities, the concept of fuzzy vectors is essential.

Definition 2: Using the notation $\mathbf{x} = (x_1, \dots, x_k)$, a k-dimensional fuzzy vector $\mathbf{x}^*$ is determined by its so-called vector-characterising function $\zeta(\cdot, \dots, \cdot)$, which is a real function of k real variables $x_1, \dots, x_k$ and has the following properties:

(1) $\zeta\colon \mathbb{R}^k \to [0,1]$

(2) The support of $\zeta(\cdot, \dots, \cdot)$ is a bounded set.

(3) For all $\delta \in (0,1]$, the δ-cut $C_\delta[\mathbf{x}^*]$, defined by

$$C_\delta[\mathbf{x}^*] \coloneqq \{\mathbf{x} \in \mathbb{R}^k \colon \zeta(\mathbf{x}) \geq \delta\},$$

is non-empty, bounded, and finite union of simply connected and closed sets.

As an example for the special case, where $k = 2$, the vector-characterising function of a measurement, or a representation of a light point on a screen, can be obtained in the following way:

Let $h(x_1, x_2)$ be the light-intensity at coordinates (x_1, x_2). The values of the vector-characterising function $\zeta(\cdot,\cdot)$ are given by

$$\zeta(x_1, x_2) \coloneqq \frac{h'(x_1, x_2)}{max\{h'(x_1, x_2)\colon (x_1, x_2) \in \mathbb{R}^2\}} \quad \forall (x_1, x_2) \in \mathbb{R}^2.$$

For higher dimensions ($k > 2$), measurements of components x_i^* are usually given by their corresponding characterising functions $\xi_i(\cdot)$. These characterising functions can be combined into a vector-characterising function $\zeta(\cdot, \dots, \cdot)$ by using a triangular norm. Especially for coordinate measurements, the product-t-norm is useful. In this case, the values of the vector-characterising function are given by

$$\zeta(x_1, \dots, x_k) = \prod_{i=1}^{k} \xi_i(x_i) \qquad \forall (x_1, \dots, x_k) \in \mathbb{R}^k.$$

Further details on characterising functions can also be found in Kovářová and Viertl (2015) and Viertl (2011).

Descriptive statistics for fuzzy data

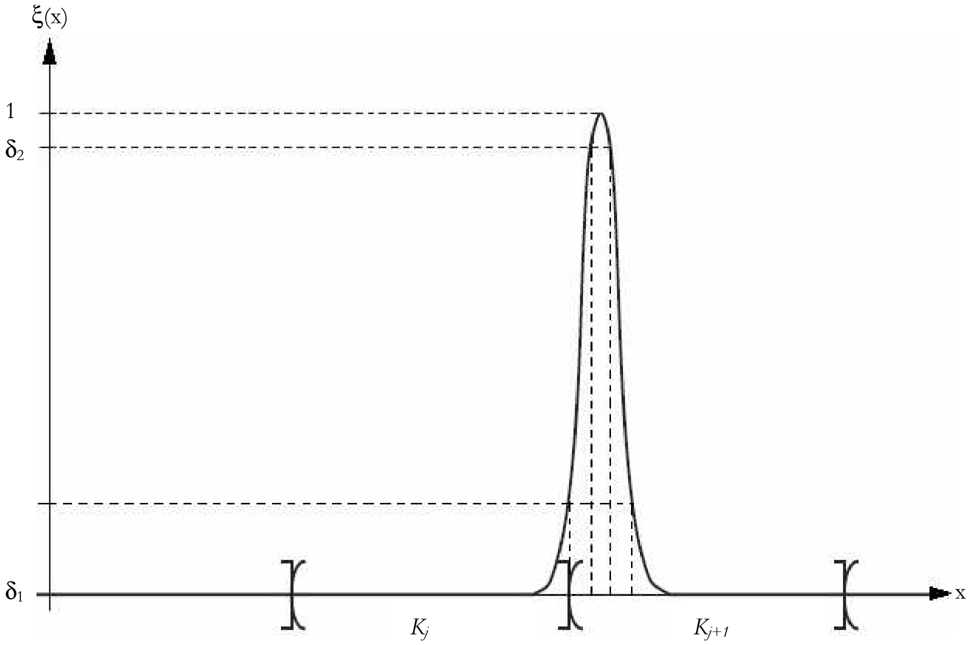
Data analysis in regional studies ranges from analysis encompassing very simple summary statistics to extremely complex multivariate analyses. This section introduces some descriptive statistics for fuzzy data with a focus on relatively simple methods. Most collected data can be used in different ways to explain the areas – variables and their behaviours –that are the main focus of the studies. The starting point for the data analysis is basic descriptive statistics, such as tables of frequencies of the main variables of interest, histograms, empirical distribution functions and correlation coefficients. This section presents these generalised descriptive statistics for fuzzy data.

Fuzzy Histograms

Given a fuzzy sample $x_1^*, \dots, x_n^*$ and a partition $K_1, \cdot\cdot\cdot, K_m$ (i.e. m classes) of the real numbers $\mathbb{R}$, the concept of relative frequencies and histograms can be extended naturally from the idealised case of real-valued samples. In this case, the most crucial aspect is that an element x_i^* may not lie within a single class but partially within different classes as depicted in Figure 1.

Figure 1

Fuzzy observation and classes of a histogram



Therefore, the relative frequency of class K_j becomes a fuzzy number $h_n^*(K_j)$. For $C_\delta(h_n^*(K_j)) = [\underline{h}_{n,\delta}(K_j), \overline{h}_{n,\delta}(K_j)]$, every δ ∈ (0,1] and every set $K_j \subseteq \mathbb{R}$ defines the lower relative frequency on a δ-cut, $\underline{h}_{n,\delta}(K_j)$, $j=1(1)m$ and the upper relative frequency of the δ-cut,$\overline{h}_{n,\delta}(K_j)$, respectively, as follows:

$$\underline{h}_{n,\delta}(K_j) := \frac{\#\{x_i^* : C_\delta(x_i^*) \subseteq K_j\}}{n}$$

$$\overline{h}_{n,\delta}(K_j) := \frac{\#\{x_i^* : C_\delta(x_i^*) \cap K_j \neq \emptyset\}}{n},$$

where # indicates cardinality, and for a fuzzy sample $x_1^*, \dots, x_n^*$, the δ-cuts are defined by $C_\delta(x_i^*) = [\underline{x}_{\delta,i}, \overline{x}_{\delta,i}]$ ∀δ∈ (0,1].

For example, given the characterising functions of a fuzzy sample of size 10 as in Figure 2, the characterising function $\eta(\cdot)$ of the fuzzy relative frequency of the class **[1, 2]** is shown in Figure 3.

Figure 2

Fuzzy sample of size 10

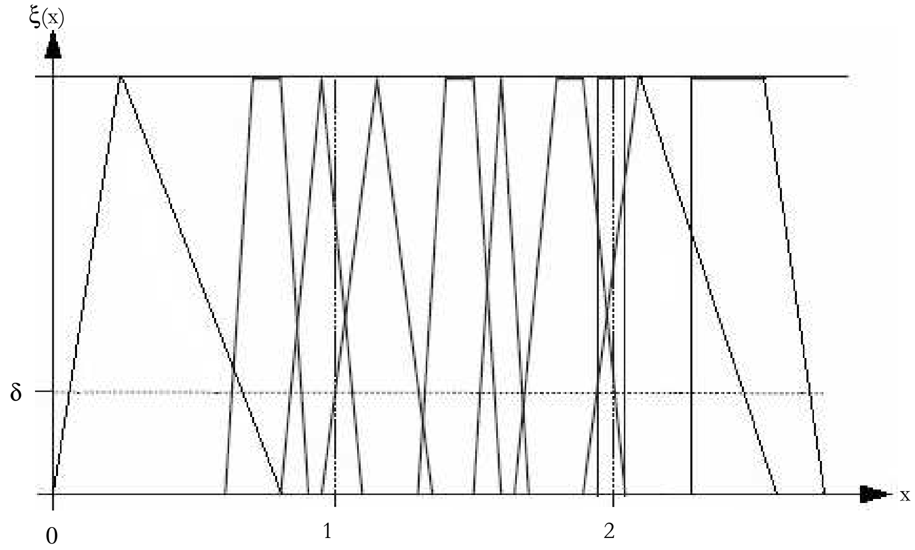
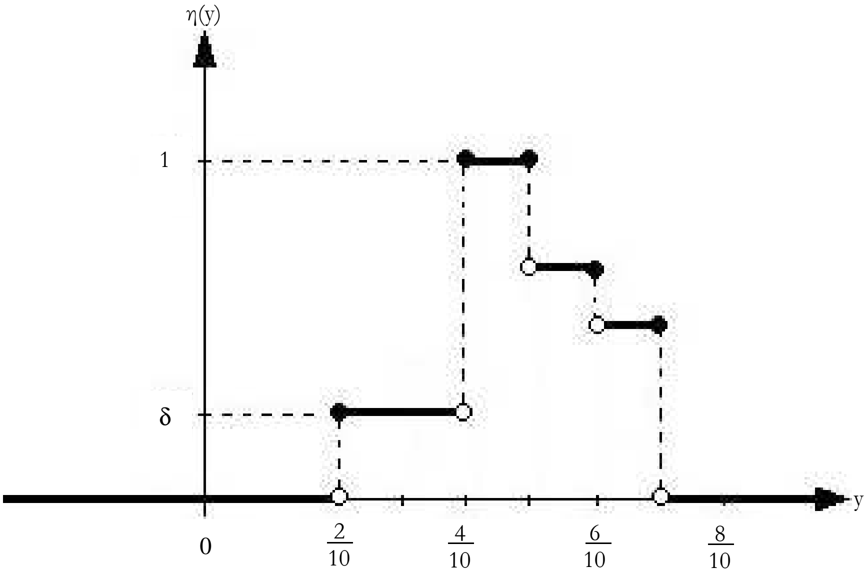


Figure 3

Characterising function of $h^*_{10}([1,2])$



In this case, at a specified δ-level, the lower (y_1 = $\underline{h}_{n,\delta}(K_j)$) and upper relative frequencies (y_2 = $\overline{h}_{n,\delta}(K_j)$) of the class [1,2], determined by $C_\delta\left(h^*_{10}([1,2])\right)$, are

$\frac{2}{10}$ and $\frac{7}{10}$ respectively. Fuzzy histograms provide more information for further statistical analysis through fuzzy probability densities.

Fuzzy empirical distribution functions

For a fuzzy sample, the empirical distribution function $\hat{F}_n^*(\cdot)$ is a fuzzy valued function defined on $\mathbb{R}$. For fixed $x \in \mathbb{R}$ and every $\delta \in (0, 1]$, $\hat{F}_{\delta,L}(\cdot)$ and $\hat{F}_{\delta,U}(\cdot)$, the lower and upper δ-level functions of $\hat{F}_n^*(\cdot)$, respectively, are defined by

$$\hat{F}_{\delta,L}(K_j) := \frac{\#\{x_i^*: C_\delta(x_i^*) \subseteq (-\infty, x]\}}{n} \text{ and}$$

$$\hat{F}_{\delta,U}(K_j) := \frac{\#\{C_\delta(x_i^*) \cap (-\infty, x] \neq \emptyset\}}{n}.$$

For a fuzzy sample $x_1^*, \dots, x_n^*$ whose δ-cuts are given by

$$C_\delta(x_i^*) = [\underline{x}_{\delta,i}, \overline{x}_{\delta,i}] \quad \forall \delta \in (0,1],$$

the corresponding δ-level functions of the fuzzy valued empirical distribution function are given by

$$\hat{F}_{\delta,L}(x) = \frac{1}{n}\sum_{i=1}^n I_{(-\infty,x]}(\overline{x}_{\delta,i}) \text{ and } \hat{F}_{\delta,U}(x) = \frac{1}{n}\sum_{i=1}^n I_{(-\infty,x]}(\underline{x}_{\delta,i}) \ \forall x \in \mathbb{R},$$

where $I_{\mathrm{A}}(\cdot)$ represents an indicator function with respect to set A.

Fuzzy correlation coefficient

For multivariate continuous data, or one observation with k variables (dimensions), idealised measurement results in a k-dimensional real vector $(x_1, \dots, x_k)$. For the special case $k = 2$, combining samples result in a fuzzy vector, denoted as $(x_1, x_2)^*$ with the vector-characterising function $\zeta(x_1, x_2) = \xi_1(x_1) \cdot \xi_2(x_2)\ \forall (x_1, x_2) \in \mathbb{R}^2$.

In case of n observations, i.e. $\mathbf{x}_i^* = (x_{1 \cdot i}, x_{2 \cdot i})^* \mathrel{\widehat{=}} \zeta_i(\cdot, \cdot)$, $i = 1(1)n$, the combined fuzzy sample $\mathbf{X}^*$is obtained by

$$\mathbf{X}^* = (x_1, x_2, \dots, x_{1 \cdot n}, x_{2 \cdot n})^* \mathrel{\widehat{=}} \zeta(x_1, x_2, \dots, x_{1 \cdot n}, x_{2 \cdot n}) \text{ , where } \zeta: \mathbb{R}^{2 \cdot n} \to [0,1].$$

In this case, the vector-characterising function $\zeta(\cdot, \dots, \cdot)$ of the combined fuzzy sample is obtained in the following way:

$$\zeta(x_1, x_2, \dots, x_{1 \cdot n}, x_{2 \cdot n}) := \min_{i=1(1)n}\{\zeta_i(x_{1 \cdot i}, x_{2 \cdot i})\}$$

$$= \min\{\zeta_1(x_1, x_2), \dots, \zeta_n(x_{1 \cdot n}, x_{2 \cdot n})\}\ \forall (x_1, x_2, \dots, x_{1 \cdot n}, x_{2 \cdot n}) \in \mathbb{R}^{2 \cdot n}$$

In other words, through the combination of n fuzzy observations $\mathbf{x}_i^*$, $i=1(1)\,n$ of a k-dimensional fuzzy quantity with vector-characterising functions $\zeta_i(\cdot)$ by the minimum-t-norm, n fuzzy k-dimensional vectors are combined into a $(k \cdot n)$-dimensional fuzzy vector with vector-characterising function $\zeta(\cdot, \dots, \cdot)$. The δ-cuts of the combined fuzzy sample $\mathbf{X}^*$ are the Cartesian products of the δ-cuts of the fuzzy vectors $\mathbf{x}_i^*$, $i=1(1)n$, which is seen from

$$\text{by } \mathsf{X}_{i=1}^n C_\delta[\xi_i(\cdot)] \Leftrightarrow \xi_i(x_i) \geq \delta\ \forall i=1(1)n \Leftrightarrow \min_{i=1(1)n}\{\xi_i(\cdot)\} \geq \delta.$$

Applying the extension principle to the following function $f_{\hat{\rho}}(\mathbf{X})$, the fuzzy sample correlation coefficient $(\hat{\rho})$ is obtained as follows:

$$f_{\hat{\rho}}(\mathbf{X}) = \frac{\sum_{i=1}^{n}(x_{1i}-\overline{x}_1)(x_{2i}-\overline{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{1i}-\overline{x}_1)^2}\sqrt{\sum_{i=1}^{n}(x_{2i}-\overline{x}_2)^2}},$$

where the characterising function of the generalised fuzzy empirical correlation coefficient r^* is given by

$$\zeta_{r^*}(r) := \left\{ \begin{array}{ll} sup\{\zeta(x_1, x_2, \dots, x_{1\cdot n}, x_{2\cdot n}) \text{: for } f_{\hat{\rho}}(\mathbf{X}) = r\} & \\ 0 \qquad \text{otherwise} & \end{array} \right\} \qquad \forall r \in [-1,1].$$

Applying the δ-cuts $C_\delta[\mathbf{X}^*]$, the lower and upper boundaries of the estimated sample correlation coefficients are obtained through simple linear programs (Shiang-Tai–Chiang 2002). A correlation coefficient provides a quantitative measure of some type of statistical relationships among the observed data values.

Models for predictions based on fuzzy information

Different methods are useful in developing models for prediction purposes, which are often of interest in regional studies, such as using historical data for projection of the next year's gross domestic product (GDP) per capita of a certain area. In general, there are two types of predictive models: parametric and non-parametric. Parametric models require some specific statistical assumptions with regard to one or more of the population parameters that characterise the underlying distribution(s), while non-parametric models are less strict with respect to the required assumptions than their parametric counterparts. The models developed for standard data have been generalised to handle fuzzy data.

Fuzzy regression

Fuzzy parametric models based on results from experiments and analysis can be constructed; for example, $\hat{Y}^* = f(x_1^*, \dots, x_k^*)$, where x_i^* are fuzzy independent variables and $\hat{Y}^*$ is a fuzzy dependent variable. In applications, there are several possibilities for taking the fuzziness into account when considering the regression models (Viertl 2011):

a) The parameters β_k and independent variables x_i are assumed to be standard real values, but the dependent variable y_i^* is fuzzy.
b) The parameters β_k^* and dependent variables y_i^* are assumed to be fuzzy, but the independent variables x_i are standard real values.
c) The dependent variables y_i as well as values of the independent variables x_i are standard real numbers, but the parameters β_k^* are fuzzy numbers.
d) The values of the independent variables x_i are fuzzy numbers x_i^*, but all other quantities are standard real numbers.

e) The independent variables and dependent variables are fuzzy, x_i^* and y_i^* respectively, but the parameters β_k are standard real values and the data set is $(x_{i1}^*, \dots, x_{ik}^*; y_i^*)$.
f) All considered quantities are fuzzy.

Frequently, quantitative regional data are collected for an analysis to model the relationship between independent and dependent variables$(x_{i1}, \dots, x_{ik}; y_i)$ for further understanding and, subsequently, for necessary prediction. Based on specific circumstances, these variables are often of uncertain (fuzzy) nature, for example, a regression model for GDP (dependent variable) with consumption, investment, and government expenditure as three independent variables. Accordingly, the fuzziness of these variables can be quantified through the methods presented in section 3. In this case, directly applied to possibility (e) for example, the independent variables and the dependent variable are fuzzy, while parameters β_k are standard real values and the data set is collected in the form of $(x_{i1}^*, \dots, x_{ik}^*; y_i^*)$. The fuzziness of these variables, without having to intuitively introduce another fuzzy coefficient into the model, can be combined before applying the extension principle. However, according to possibility (b), as originally proposed by Tanaka et al. (1982) and the most frequently used method, fuzzy regression models assume a fuzzy dependent variable and a fuzzy coefficient, but crisp independent variables to minimise the fuzziness of the model (Shapiro 2006).

In case of a k-dimensional sample of observations $\mathbf{x}_i^*$, i=1(1) n, the generalised minimum rule is applied to obtain the vector-characterising function $\zeta(x_1, \dots, x_{kn})$ for the combined fuzzy vector $\mathbf{X}^*$, which is the combined fuzzy sample.

Considering a fuzzy sample $\mathbf{x}_1^*, \dots, \mathbf{x}_n^*$ with the corresponding vector-characterising functions $\zeta_i(\cdot, \dots, \cdot)$, where $\mathbf{x} = (x_1, \dots, x_k) \in \mathbb{R}^k$ and sample space $\mathbb{R}^{k\,n}$, that is $\mathbf{x}_i^* \in \mathcal{F}(\mathbb{R}^n)$, i=1(1) n and the combined fuzzy sample $\mathbf{X}^* \in \mathcal{F}(\mathbb{R}^{k\,n})$, through the combination of n fuzzy observations $\mathbf{x}_i^*$, i=1(1) n of a k-dimensional fuzzy quantity by the minimum-t-norm, n fuzzy k-dimensional vectors are combined into an $(k \cdot n)$-dimensional fuzzy vector with vector-characterising function $\zeta(\cdot, \dots, \cdot)$, for which the following property holds:

$$C_\delta[\zeta(\cdot, \dots, \cdot)] = \mathrm{X}_{i=1}^{n} C_\delta\left[\zeta_i(\cdot)\right] \qquad \forall\, \delta \in (0,1],$$

where $\zeta(x_1, \dots, x_{k \cdot n})$=min$\{\zeta_1(x_1, \dots, x_k), \zeta_2(x_{k+1}, \dots, x_{2k}), \dots,$
$\zeta_n(x_{(n-1)k+1}, \dots, x_{kn})\}\quad \forall (x_1, \dots, x_{k\,n}) \in \mathbb{R}^{kn}$

Let η_i, for i = 1(1)n, denote the characterising function of y_i^* and the combined fuzziness is contained in the fuzzy element $\mathbf{t}^*$of $\mathbb{R}^{(k+1)n}$, whose vector-characterising function $\tau(\cdot, \dots, \cdot)$ is defined by

$$\tau(\mathbf{x}_1, \dots, \mathbf{x}_n, y_1, \dots, y_n) := \min\{\zeta_1(\mathbf{x}_1), \dots, \zeta_n(\mathbf{x}_n), \eta_1(y_1), \dots, \eta_n(y_n)\}$$

$$\forall \begin{cases} x_i \in \mathbb{R}^k \\ y_i \in \mathbb{R} \end{cases}.$$

Based on this fuzzy element $\mathbf{t}^*$ in $\mathbb{R}^{(k+1)n}$, the estimators $\hat{\beta}_k$ for the regression parameters can be generalised. The characterising function $\phi_j(\cdot)$ of the fuzzy estimator $\hat{\beta}_j^* \in \mathbb{R}$ is given by

$$\phi_j(z) := \begin{cases} sup\{\tau(\mathbf{t}): \hat{\beta}_k(\mathbf{t}) = z\} & \text{if } \exists \mathbf{t} \in \mathbb{R}^{(k+1)n}: \hat{\beta}_k(\mathbf{t}) = z \\ 0 & \text{if } \nexists \mathbf{t} \in \mathbb{R}^{(k+1)n}: \hat{\beta}_k(\mathbf{t}) = z \end{cases} \quad \forall z \in \mathbb{R}.$$

A generalised least-squares method may be used in approximating the crisp regression coefficients $\hat{\beta}_k$. The estimated fuzzy regression model can be built as

$$\hat{Y}_i^* = \hat{\beta}_0 \oplus \hat{\beta}_1 \odot x_{i1}^* \oplus \hat{\beta}_2 \odot x_{i2}^* \oplus \ldots \oplus \hat{\beta}_k \odot x_{ik}^*,\ i = 1(1)n,$$

where $\hat{Y}_i^*$ and x_{ik}^* represent the estimated fuzzy dependent variable and the kth fuzzy independent variable of the i-th observation, respectively. The predictions of dependent values from a specified model are results from applying the generalised algebraic operations (multiplications and additions) for fuzzy quantities.

Fuzzy time series

The main objective of time series analysis is to build mathematical models based on known trends and seasonal influences from historical data for future prediction. A fuzzy time series x_t^*, where $t \in T = \{1,2,3,\ldots, N\}$, is an ordered sequence of fuzzy numbers. In other words, a one-dimensional fuzzy time series is a mapping $T \rightarrow \mathcal{F}(\mathbb{R})$, which results in a fuzzy number x_t^* at any time point t. Different descriptive methods of fuzzy time series analysis, such as moving averages (a filtering method), have been well developed.

Moving averages apply the concepts of local approximation through a local arithmetic mean to eliminate the random oscillations of observed fuzzy time series data x_t^*. Through the extension principle, the fuzzy numbers $x_{t-q}^*, \ldots, x_{t+q}^*$ can be combined into a fuzzy vector $\mathbf{x}^* \in \mathcal{F}(\mathbb{R}^{2q+1})$, which is determined by its vector-characterising function $\zeta_{\mathbf{x}^*}(\cdot, \ldots, \cdot)$. As a result, the smoothed time series y_t^* is obtained, where $t = q+1(1)N-q$ and q denotes the length of moving averages. The characterising function of y_t^* and the δ-cut $C_\delta[y_t^*]$ of the smoothing through a local arithmetic mean are derived as in (Viertl 2011) by

$$\zeta_{y_t^*}(y) = sup\left\{\zeta_{\mathbf{x}^*}(x_{-q}, \ldots, x_q): (x_{-q}, \ldots, x_q) \in \mathbb{R}^{2q+1} : \frac{1}{2q+1}\sum_{i=-q}^{q} x_i = y\right\},$$

and

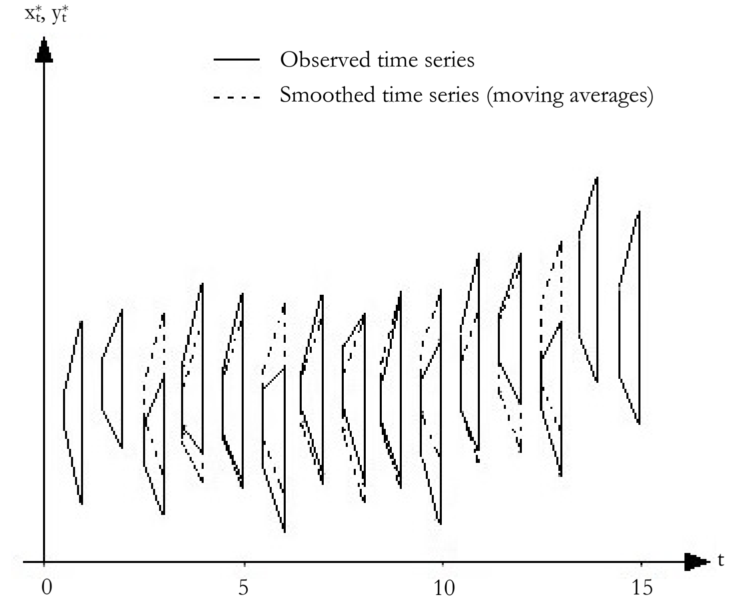
$$C_\delta(y_t^*) = \left[\min_{(x_{-q},\ldots, x_q)\in C_\delta(\mathbf{x}^*)} \frac{1}{2q+1}\sum_{i=-q}^{q} x_i\,, \max_{(x_{-q},\ldots, x_q)\in C_\delta(\mathbf{x}^*)} \frac{1}{2q+1}\sum_{i=-q}^{q} x_i\right],$$

respectively.

As an example, let $T = \{1,2,3,\ldots,15\}$ and a fuzzy time series x_t^* with trapezoidal characterising functions, that $(x_t^*)_{t=1(1)15} = \left(t^*(m_t, s_t, l_t, r_t)\right)_{t=1(1)15}$. As results of applying moving averages of length 2, the characterising functions of the values of the smoothed time series $(y_t^*)_{t=3(1)13}$ are shown (trapezoids with dashed line) in Figure 4.

Figure 4

Fuzzy time series with moving averages



The filtered time series is smoother if more observations are considered for the filtration, i.e. larger q, with exceptions on the boundaries where the filtered values cannot be obtained. On the other hand, the smoothed time series (y_t^*) is shorter than the original time series (x_t^*).

Fuzzy predictive densities

In Bayesian inference, probabilities of events $A\in\mathcal{A}$ based on a fuzzy probability density $f^*(\cdot)$ are relevant. The standard predictive density for stochastic model $X\sim f(\cdot\mid\theta)$; $\theta\in\Theta$, based on data D, is defined as marginal density of the stochastic quantity X of $(X,\tilde{\theta})$, that is the following:

$$p(x|D) \coloneqq \int_{\Theta} f(x|\theta)\pi(\theta|D)\,d\theta \quad \forall x \in M_x,$$

where M_x is the observation space of X and $\pi(\cdot\,|D)$ is the a-posteriori density of the parameter. Fuzzy probability densities are a more general form of expressing a-priori information concerning the parameters θ in stochastic models $f(\cdot\mid\theta)$, $\theta\in\Theta$. The generalisation of the predictive density for fuzzy a-posteriori densities $\pi^*(\cdot\mid D^*)$ based on fuzzy data D^* is possible in the following way (Viertl–Sunanta 2013):

Let D_δ be the set of all standard probability densities $h(\cdot)$ on Θ with $\underline{\pi}_\delta(\theta \mid D^*) \leq h(\theta) \leq \overline{\pi}_\delta(\theta \mid D^*)$ $\theta \in \Theta$; $\underline{\pi}_\delta(\theta)$ and $\overline{\pi}_\delta(\theta)$ are lower and upper bounds of the densities at each δ-level.

In case of fuzzy a-posteriori density $\pi^*(\cdot \mid D^*)$, the integration has to be generalised accordingly. This generalised integration yields fuzzy intervals as a result. Based on D_δ, the generating family of intervals $[a_\delta, b_\delta]$, $\delta \in (0,1]$ is defined by

$a_\delta := \inf\{\int_\Theta f(x|\theta)h(\theta)d\theta : h \in D_\delta\}$

$b_\delta := \sup\{\int_\Theta f(x|\theta)h(\theta)d\theta : h \in D_\delta\}$.

Definition 3: The fuzzy predictive density $p^*(\cdot \mid D^*)$ is defined by its values $y^* = p^*(x \mid D^*)$ $\forall x \in M_x$, whose characterising function $\psi_x(\cdot)$ is given, through the construction lemma (see also Viertl 2011), by

$\psi_x(y) = \sup\{\delta . \mathbb{1}_{[a_\delta, b_\delta]}(y) : \delta \in [0,1]\}$ $\forall y \in \mathbb{R}$, where $[a_0, b_0] := \mathbb{R}$.

In other words, the fuzzy value of the generalised predictive density is defined via the family of nested compact intervals $[a_\delta, b_\delta]$. The fuzzy predictive distribution is used in making probabilistic statements of the unobserved x without explicit conditioning on parameters θ, but with conditioning on previously collected fuzzy data D^*.

Fuzzy data in databases

As part of building a complete system, data and information are obtained, analysed, and stored in databases. The important information generally comes from different sources and often cannot be replicated, such as estimation from human experts who describe their knowledge about the areas of interest in natural languages, sensory measurements and mathematical models derived according to physical laws with respect to the systems of interest. Many practical applications require data management components that provide support for managing uncertain data. There are different types of uncertain data: imprecise, vague, ambiguous, inconsistent and incomplete data (Popat–Sharda–Taniar 2004). Fuzzy theory allows us to develop models for imprecise or vague data, in other words, to integrate the vague knowledge into databases. To store this type of information, fuzzy databases are necessary for storing fuzzy data.

There are several efforts for extending relational database systems in order to represent imprecise data and queries. For example, the work by Serrano et al. (2001) shows that fuzzy models can work with the imprecision and uncertainty associated with agriculture information in relational databases. The fuzzy relation and fuzzy set theory provide a requisite mathematical framework for dealing with such fuzzy data (Guglani–Katti–Saxena 2013). Fuzzy relational database theory extends the relational model to allow for the representation of imprecise data and, thus, provides a more accurate representation of the intended information. In other words, applied databases must be able to store fuzzy numbers and fuzzy vectors in order to provide realistic information concerning real data. Fuzzy numbers and fuzzy vectors can be represented in databases by storing δ-cuts. In addition, fuzzy multivariate data can be represented in databases by storing a suitable family of δ-cuts of the corresponding vector-characterising function. Learning how to store fuzzy data in traditional relational databases is critical to satisfying the normal forms and keeping the integrity of a database through the fuzzy meta-model of a relational database. A fuzzy meta-model keeps all relevant fuzzy data and manages links to the relations of real entities (see also Hudec 2016 for details).

Final remarks

In regional studies, measurements are crucial in data collection for further statistical analyses. These measurements of continuous variables are uncertain, or more or less fuzzy. The fuzziness of individual measurement results can be described by so-called fuzzy numbers, whereas the variability and errors are described by stochastic models. As a result, the analysis of repeated measurements is possible using respective generalised statistical methods. In this contribution, some generalised statistical methods to handle the so-called fuzzy data are described. Descriptive statistics provide simple summaries of the collected samples and measures (data). They form the basis of virtually every quantitative analysis of the data. Through concepts of fuzzy numbers and characterising functions, fuzzy data are summarised and represented in forms of fuzzy histograms, which provide more information when memberships of the individual data to different classes are not crisp. Some other statistics, such as fuzzy empirical distribution functions and correlation coefficients, are also useful for preliminary data analysis. For better understanding and future projection of the behaviours of the variables under analysis, models for prediction based on fuzzy information, such as fuzzy regression, fuzzy time series and fuzzy predictive density, have been generalised and introduced.

Fuzziness is everywhere in the physical world. In order to describe different regional facets of reality, the methods have to undertake this type of uncertainty. This is possible, and related methods are available through mathematical models. Accordingly, application of such methods results in more realistic models for data analysis and, subsequently, better understanding of the collected data.

REFERENCES

AMT DER TIROLER LANDESREGIERUNG (2016): *Stickstoffdioxid: Grenz- und Richtwerte, Abteilung Raumordnung-Statistik* (downloaded: 8 June 2016) https://www.tirol.gv.at/umwelt/luft/ diagramm-stickstoffdioxid/

BURROUGH, P. A. (2001): GIS and Geostatistics: Essential Partners for Spatial Analysis *Environmental and Ecological Statistics* 8 (4): 361–377.

ENE, C. M.–HURDUC, N. (2010): A fuzzy Model to Estimate Romanian Underground Economy *Internal Auditing and Risk Management* 2 (18): 1–10.

FEDERAL ENVIRONMENT AGENCY (2002): *6th Report on the State of the Environment in Austria* Wien. (downloaded 8th June 2016) http://www.umweltbundesamt.at/fileadmin/site/umweltkontrolle/2001/E-02_luft.pdf

GUGLANI, S.–KATTI, C.P.–SAXENA, P. C. (2013): Fuzzy Statistical Database and Its Physical Organization *International Journal of Database Management Systems* 5 (4): 27–47.

GULEDA, O. E.–IBRAHIM, D.–HALIL, H. (2004): Assessment of Urban Air Quality in Istanbul using Fuzzy Synthetic Evaluation *Atmospheric Environment* 38 (23): 3809–3815.

HUDEC, M. (2016): *Fuzziness in Information Systems* Springer, Switzerland.

KLIR, G.–YUAN, B. (1995): *Fuzzy Sets and Fuzzy Logic-Theory and Applications* Prentice Hall, Upper Saddle River.

KOVÁŘOVÁ, L.–VIERTL, R. (2015): The Generation of Fuzzy Sets and the Construction of Characterizing Functions of Fuzzy Data *Iranian Journal of Fuzzy Systems* 12 (6): 1–16.

LANDESINSTITUT FÜR STATISTIK (ASTAT) (2015): *Statistisches Jahrbuch für Südtirol Autonome Provinz Bozen*, Bozen, Südtirol.

LEE, E. S. (1995): Fuzzy Spatial Statistics In: *Selected Papers of Engineering Chemistry and Metallurgy* pp. 151–157., Institute of Chemical Metallurgy, Chinese Academy of Science, China.

POPAT, D.–SHARDA, H.–TANIAR, D. (2004) Classification of Fuzzy Data in Database Management System In: NEGOITA, M. G. (Ed.*) Proceedings of Knowledge-Based Intelligent Information and Engineering Systems* pp. 691–697., 8th International Conference, New Zealand.

SERRANO, J. M.–VILA, M. A.–ARANDA, V.–DELGADO, G. (2001): Using Fuzzy Relational Databases to Represent Agricultural and Environmental Information *Mathware & Soft Computing* 8: 275–289.

SHAPIRO, A. F. (2006): *Fuzzy Regression Models* In: Proceedings of Actuaries Research Conference (ARC), Instituto Tecnológico Autónomo de México (ITAM), Mexico, August 11-13, 2005, Society of Actuaries, IL.

SHIANG-TAI, L.–CHIANG, K. (2002): Fuzzy Measures for Correlation Coefficient of Fuzzy Numbers *Fuzzy Sets and Systems* 128 (2): 267–275.

TANAKA, H.–UEJIMA, S.–ASAI, K. (1982): Linear Regression Analysis with Fuzzy Model *IEEE Transactions on Systems, Man and Cybernetics* 12 (6): 903–907.

VIERTL, R. (2011): *Statistical Methods for Fuzzy Data* Wiley, Chichester.

VIERTL, R. (2015): Measurement of Continuous Quantities and their Statistical Evaluation *Austrian Journal of Statistics* 44 (1): 25–32.

VIERTL, R.–SUNANTA, O. (2013): Fuzzy Bayesian Inference *METRON (Fuzzy Statistical Analysis: methods and applications)* 71 (3): 207–216.

WICHERN, D. W.–JOHNSON, R. A. (2007): *Applied Multivariate Statistical Analysis* 6th ed., Pearson Prentice Hall, NJ.