Statistical Analysis of Uncertain Measurements in Modeling of Water and Wastewater Systems

Olgierd Hryniewicz, Jan Studzinski

Abstract
Statistical data used in the analysis of environmental problems are very often presented in an imprecise way. We make a distinction between random uncertainty that can be described by stochastic models and linguistic lack of precision that may be modelled by fuzzy sets or possibilistic distributions. In the paper we present a short introduction to fuzzy statistics which may be useful for the analysis of such kind of data. This methodology is illustrated with the example of the communal water supply prediction problem when the statistical data include both the results of measurements and imprecise expert opinions.

1. Introduction
Statistical methods have been widely used in all areas related to environmental sciences. A collection of statistical methods that are typical for solving problems encountered in environmental sciences has been given a specific name – environmetrics. This special name makes it clear that these methods are sometimes somewhat different from the methods used in other fields of application. A good description of the most frequently used methods is given, for example, in (Barnett 2004). A much more comprehensive description of statistical methods used for solving environmental problems can be found in (El-Shaarawi/Piegorsch 2002). However, what does not distinguish environmetrics from other fields of statistics is its usage of purely stochastic models for the description of uncertainties of all kind.

Statistical data analysed while solving environmental problems may be of a different kind. If the uncertainty of statistical data is purely of a random character the usage of classical statistical methods does not raise any questions. However, in many cases statistical data is not only random but also imprecise. This happens especially often when the data contains information which is reported by humans using an imprecise plain language. Consider, for example, the information given by an expert that a certain parameter has a value "about five". For many years such imprecise information has been modelled by probability distributions. However, probabilistic description of the so called "linguistic uncertainty" has raised many questions. Some authors, especially from the area of computer data analysis (data mining, knowledge discovery) have claimed that other theories, such as fuzzy set theory and possibility theory, are more suitable for the description of uncertainties of that type. It must be stressed however, that the question mentioned above has not been answered yet. For more detailed discussion of the problem the readers are encouraged to read a special issue of the journal "Reliability Engineering and System Safety", and its editorial (Helton/Oberkampf 2004).

In this paper we present the methodology which can be useful for solving statistical problems where both type of uncertainty, random and linguistic, are present in the data. Following (Grzegorzewski/Hryniewicz 1999), in the second section we introduce the notion of a fuzzy random variable that will be used for the formal description of imprecise random data. We recall some basic notions of fuzzy statistics, and discuss the most important features of the proposed approach. In the third section of the pa-
per we will present a rather simple example of dealing with imprecise statistical data which is used for making decisions in the design of a communal water supply system.

2. Fuzzy random variables and fuzzy statistics

Random variable is the basic notion used in the description of random phenomena. It is a mapping which assigns to each random event a real number. However, when the available data are characterized by a non-random (linguistic) imprecision we need another formalism to describe our data. If the data consists of imprecisely reported numbers we can use the formalism of the fuzzy sets theory which is the generalization of the classical set theory. In the fuzzy sets theory the imprecise counterpart of a real number is a fuzzy number defined by its membership function. Formally, a fuzzy number can be defined as follows.

Definition 1 (Dubois/Prade 1983)
The fuzzy subset $A$ of the real line $R$, with the membership function $\mu : R \rightarrow [0,1]$, is a fuzzy number if

- is normal, i.e. there exists an element $x_0 \in R$ such that $\mu(x_0) = 1$;
- is fuzzy convex, i.e. $\mu(\lambda x + (1-\lambda)y) \geq \mu(x) \wedge \mu(y)$ for all $x,y \in R$ and $0 \leq \lambda \leq 1$;
- is upper semicontinuous;
- $\supp(\mu)$ is bounded.

A useful concept used for the description of fuzzy numbers is the $\alpha$-cut. The $\alpha$-cut $A_\alpha$ of a fuzzy number $A$ is a non-fuzzy set defined as

$$A_\alpha = \{x \in R: \mu(x) \geq \alpha\}.$$

The family $\{A_\alpha : \alpha \in [0,1]\}$ is a set representation of the fuzzy number $A$. Basing on the resolution identity, we have the alternative description of fuzzy numbers:

$$\mu(x) = \sup_{\alpha \in [0,1]} \{\alpha I_{A_\alpha}(x)\},$$

where $I_{A_\alpha}(x)$ denotes the characteristic function of $A_\alpha$. Definition 1 implies that every $\alpha$-cut of a fuzzy number is a closed interval. Hence, we have $A_\alpha = [L_{\alpha}, U_{\alpha}]$, where

$$L_{\alpha} = \inf \{x \in R: \mu(x) \geq \alpha\},$$

$$U_{\alpha} = \sup \{x \in R: \mu(x) \geq \alpha\}.$$

The space of all fuzzy numbers will be denoted by $F(R)$.

In case we deal with uncertainties of both, random and linguistic, types we can use the notion of a fuzzy random variable. A fuzzy random variable may be defined by analogy to the definition of a real-valued random variable as a mapping that assigns to a random event an imprecise fuzzy number. The notion of a fuzzy random variable has been defined independently by many authors. The definition given below is taken from (Grzegorzewski/Hryniewicz 1999), and is similar to original definitions of (Kwakernaak 1978) and (Kruse 1982). Suppose a random experiment is described as usual by a probability space $(\Omega, F, P)$, where $\Omega$ is the set of all possible outcomes of the experiment, $F$ is a $\sigma$-algebra of subsets of $\Omega$ (the set of all possible events) and the function $P$, defined on $F$, is a probability measure.

Definition 2
A mapping $X : \Omega \rightarrow F(R)$ is called a fuzzy random variable if it satisfies the following properties:

1. $\{X_\alpha(\omega) : \alpha \in [0,1]\}$ is a set representation of $X(\omega)$ for all $\omega \in \Omega$,
2. for each $\alpha \in [0,1]$ both $X_{\alpha}^L$ and $X_{\alpha}^U$ defined as
are real-valued random variables on \((\Omega, \mathcal{F}, P)\). Thus, a fuzzy random variable \(X\) is considered as a perception of an unknown usual random variable \(V: \Omega \to \mathbb{R}\), called an original of \(X\). Let \(\chi\) denotes a set of all possible originals of \(X\). If only vague data are available, it is of course impossible to show which of the possible originals is true. Therefore, we can define a fuzzy set of \(\chi\), with a membership function \(\nu: \chi \to F(\mathbb{R})\) given as follows:

\[
\nu(V) = \inf \{ \mu_X(\omega)(V(\omega)) : \omega \in \Omega \}
\]

which corresponds to the grade of acceptability that a fixed random variable \(V\) is the original of the fuzzy random variable in question.

Fuzzy random variables have been used for the description of many practical problems where stochastic randomness is present together with fuzzy imprecision. Classical statistical methods have been also generalized to the case of the analysis of fuzzy random data. The readers are encouraged to read the book (Kruse/Meyer 1987), recent collections of papers in (Grzegorzewski 2002) and (Lopez-Diaz 2004), and numerous papers published in many journals, such as, for example, *Fuzzy Sets and Systems*.

One of the most frequently used methods in the statistical analysis of fuzzy data is the estimation of the fuzzy parameters of the probability distribution. A general theory of this procedure can be found in (Kruse/Meyer 1987). In the next section we present a simple example how this methodology could be used for forecasting future consumption of communal water.

3. **Forecasting of future water consumption using imprecise statistical data and experts opinions**

To illustrate the usage of the fuzzy statistics let us consider the example of a future water consumption forecast. Suppose that a new residential area of a city is going to be built in a five years period. In order to design a communal water supply system it is necessary to evaluate a probability distribution for future one-day water consumption in this area in the time period characterised by the highest water consumption.

The process of water consumption of an individual household is usually a rather complicated non-stationary random process. Moreover, it is very difficult to forecast its future value, as it heavily depends on technologies of future home appliances, future social trends related to environmental behaviour, etc. There also exists significant uncertainty as to the future number of households in the considered area. To put it in a nutshell: building a mathematical model allowing forecasting of future water consumption for this residential area is, in principle, a very hard task. This task is even harder, if at all possible, when we have to identify the model’s parameters using the existing statistical data. Therefore, there is a need to build a simpler model that would be useful for merging information from imprecise statistical data and also imprecise opinions of experts.

Suppose, that the only available source of statistical data is a data-base with records of approximately monthly (or even bi-monthly) water consumption of households that could be consider similar to the households in the new residential area. Let \(n\) be the number of analysed households, and for each household we observe pairs of numbers \((v, t)\), where \(v\) is the water consumption during the period of length \(t\) (to simplify the notation we suppressed here the time index). To simplify further the model let us assume, that the season of the highest consumption is already known, so for the future analysis we use only one data point \((v_i, t_i), i = 1, \ldots, n\) for each analysed household (Note, that the analysed time periods may overlap, and may have different lengths). Having these data we can calculate the current average consumption per day from a simple formula

\[
X^L_\alpha = X^H_\alpha(\omega) = \inf X^L_\alpha,
\]

\[
X^U_\alpha = X^H_\alpha(\omega) = \sup X^L_\alpha,
\]
Knowing these values we can easily calculate such statistical characteristics as the average
\[ \bar{c} = \frac{1}{n} \sum_{i=1}^{n} c_i \]
and standard deviation
\[ \sigma_c = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (c_i - \bar{c})^2} \]
However, it is easy to notice that these values are very rough estimates of the parameters of the probability distribution of one-day consumption, as their input values represent averaged values over a usually long period of time. Therefore, they seem to be not sufficient for the evaluation of the probability distribution of the future one-day water consumption.

Now, let us assume that we have opinions of \( m \) experts who are asked about the values of coefficients \( b \) that define the relation between the future one-day consumption \( x \) and its current value \( c \). We assume the simplest possible relation, i.e. \( x = bc \).

Let the experts, whom we ask about the value of \( b \), provide three values: minimal possible value \( b_{\text{min}} \), the most possible value \( b_0 \), and maximal possible value \( b_{\text{max}} \). Such three values, in a general case denoted below by a fuzzy variable \( z \), may be used for the construction of the possibility distribution of the imprecisely defined fuzzy triangular number \( (z_{\text{min}}, z_0, z_{\text{max}}) \) with the following membership function:
\[
\mu_T(z) = \begin{cases} 
0 & z < z_{\text{min}} \\
\frac{z - z_{\text{min}}}{z_0 - z_{\text{min}}} & z_{\text{min}} \leq z < z_0 \\
1 & z = z_0 \\
\frac{z_{\text{max}} - z}{z_{\text{max}} - z_0} & z_0 \leq z \leq z_{\text{max}} \\
0 & z > z_{\text{max}} 
\end{cases}
\]
If only one value of a fuzzy coefficient \( \tilde{b} \) is needed we can find it by averaging \( m \) triangular fuzzy numbers \( (b_{\text{min},j}, b_0, b_{\text{max},j}) \) using the Zadeh's extension principle.

**Definition 3. Extension principle (Dubois/Prade 1980)**

Let \( X \) be a Cartesian product of universe \( X = X_1 \times X_2 \times \cdots \times X_r \), and \( A_1, \ldots, A_r \) be \( r \) fuzzy sets in \( X_1, \ldots, X_r \), respectively. Let \( f \) be a mapping from \( X = X_1 \times X_2 \times \cdots \times X_r \) to a universe \( Y \) such that \( y = f(x_1, \ldots, x_r) \). The extension principle allows us to induce from \( r \) fuzzy sets \( A_i \) a fuzzy set \( B \) on \( Y \) through \( f \) such that
\[
\mu_B(y) = \sup_{x_1, \ldots, x_r; y = f(x_1, \ldots, x_r)} \min[\mu_{A_1}(x_1), \ldots, \mu_{A_r}(x_r)]
\]
\[
\mu_B(y) = 0 \text{ if } f^{-1}(y) = \emptyset
\]
Using the extension principle we can easily show that the average fuzzy coefficient \( \tilde{b} \) is also a triangular fuzzy number defined as \( (b_{\text{min}}, b_0, b_{\text{max}}) \), where
\[
b_{\text{min}} = \frac{1}{m} \sum_{j=1}^{m} b_{\text{min},j}, \quad b_0 = \frac{1}{m} \sum_{j=1}^{m} b_0, j, \quad b_{\text{max}} = \frac{1}{m} \sum_{j=1}^{m} b_{\text{max},j}.
\]
Now, let us suppose that the future number of households in the considered residential area is also imprecisely evaluated by an expert, and expressed by a triangular fuzzy number $\tilde{N}$ defined by three real numbers $(N_{\min}, N_0, N_{\max})$. In such a case we can find that the fuzzy estimator of the expected value of the future one-day water consumption in the considered residential area is given by a fuzzy triangular number

$$(N_{\min}b_{\min}, N_0b_0, N_{\max}b_{\max})$$

and the fuzzy estimator of its standard deviation is given by a fuzzy triangular number

$$(N_{\min}b_{\min}\sigma_c, N_0b_0\sigma_c, N_{\max}b_{\max}\sigma_c).$$

If we assume that the one-day consumption from a household is described by a normal distribution, these two fuzzy triangular numbers define a family of fuzzy normal distributions for future one-day water consumption in the considered residential area. It is worth noting, that this description let us incorporate different types of uncertainties: one that arose from random statistical data and second that was due to imprecisely expressed experts opinions.

The example considered in this section is given as a simple illustration of a really existing problem. In real life not all simplifications described above can be applicable. For example, in a more realistic setting it could be necessary to use different values of coefficients $b$ for different subgroups of households. In such a case, the fuzzy standard deviation of the future one-day water consumption will be no longer described by a fuzzy triangular number, making all computations more cumbersome. Nevertheless, the proposed methodology will be still valid. Moreover, it will be also possible to use the methods of fuzzy statistics for building fuzzy confidence intervals, and for testing statistical hypotheses.

Mathematical formalism proposed in this paper will be tested in solving practical problems of the water consumption forecast for a water supply system in one of Polish cities. The project has been already started at the Systems Research Institute in cooperation with a commercial partner.

**Bibliography**


