STATISTICAL SOFTWARE FOR RAW WATER QUALITY ASSESMENT

M. Milivojevic[1], Dj. Forst[1], Dj. Moljkovic[1], M. Tomic[1]
[1]Technical and Business College, Uzice, Serbia,
{milovan.milivojevic, djordje.forst}@vpts.edu.rs, djmoljkovic@gmail.com, theincrediblestark@gmail.com

Abstract: *Based on the importance of drinking water in the 21st century and insight into leading edge trends in the domain of management of drinking water (number of scientific publications, statistical and software support), this paper presents AQUA Statistic software for statistical evaluation of raw water quality, which the authors of this paper developed themselves, using the C# 6.0 programming environment. Special emphasis is placed on software modules dealing with analysis of variance (One-way ANOVA, Two-way ANOVA). Modules of AQUA Statistic software are validated for the example of raw water electrical conductivity in the Case Study and on the dataset of raw water properties, collected in the district of Zlatibor in the southwest part of the Republic of Serbia. AQUA Statistic software has the ability to automate the Integrated, as well as the ability to incorporate numerous Artificial Intelligence and Data Mining algorithms based on open sources platforms, such as R and Python.*

Keywords: *Quality of Raw Water, ANOVA, Statistical software.*

1. INTRODUCTION AND STATE OF THE ART

The time of accelerated and intense changes in the field of science and applied technology in the second decade of the 21st century is characterized by topics and problems related to sustainable drinking water resources on planet Earth. Massive industrial production, intensive food production, pesticide application, herbicide and fungicide in crop protection, climate change, concentrated pollution in big cities, nuclear weapons trials, regional conflicts and wars, exponential increase in population ... are just some of the entries that determine this essential problem of the sustainable development of civilization. Drinking water becomes *the gold* of the 21st century.

For responsible drinking water management, as a limited resource, at local and global level, national policies and sustainable development strategies must be based on the application of all available methods of modern science and their implementation based on modern sensor and measuring equipment. Therefore, in recent years, an extremely large number of scientific papers dealing with this domain have been published. However, although AI techniques and Data Mining paradigms are increasingly being applied in modeling quality, processes and properties related to raw and drinking water, the traditional statistical approach is still relevant. For example, in [1], Ammar T. A. et al. investigate the impact of chlorine dioxide as one of the most promising in water treatment. This work provides a novel mathematical equation for chlorine dioxide decay prediction in desalinated water. To confirm the validity of the proposed decay rate model/equation, site verification was performed (real concentration vs. predicted concentration) and then t-test formula was used to indicate the similarity of both test results. Beaudeau P. et al. in [2] used a Poisson regression to compare daily hospital admissions of elderly people for acute gastrointestinal illness in Boston against daily variations in drinking water quality over an 11-year period, controlling for weather, seasonality and time trends. Water quality data included turbidity, fecal coliforms, UV-absorbance, and planktonic algae and cyanobacteriae concentrations. In study [3], authors presented a one-year sampling program which covering twenty-five small municipal systems was carried out in two Canadian regions to improve understanding of the variability of water quality in small systems from water source. In order to determine the most important parameters for explaining the spatio-temporal variability of chlorinated disinfection by-products and free residual chlorine, stepwise analysis was applied. These compounds have been under study for several years, and epidemiological and toxicological studies have suggested potential negative effects on human health. In paper [4], Dahhoua M. et al. use a statistical approach to evaluate the degree of metal pollution, trace element concentrations, and seasonal evolutions of various physicochemical parameters (volatile suspended solids, suspended matter, conductivity, pH(s), the content of element such as e.g.: Pb, Cr, Cd, Fe, Al, Cu, Zn, P, N, K etc) of Moroccan drinking water sludge and in dried hydroxide sludge. In study [5], the data of nine water quality variables (T, ECw, DO, SO_4^{2-}, Na^++K^+, Mg^{2+}, Ca^{2+}, NO_3^-, TP) in the Strymon river of Greece for the period 1980-1997 were selected for analysis. Time series were analyzed and additional χ^2-test and the Kolmogorov-Smirnov test were used to select the theoretical distribution which best fitted the data. Trends were detected using the nonparametric Spearman's criterion. In [6], the concentration and spatial distribution of nitrate in the Merida's karstic aquifer (Merida city, Mexico,) were assessed by statistical and geostatistical techniques, beacause water containing nitrate levels above 45 mg/l is not recommended for human consumption and its prolonged intake is associated with various health conditions. Non-parametric methods were used to prove the hypotheses of evenness among temporal-spatial evaluation for water supply systems given that raw data did not followed a normal distribution. The Kolmogorov–Smirnov test was applied to prove the temporal evenness hypothesis. Based on this result, the nitrate concentration analysis was performed using a non-parametric Kruskal–Wallis Analysis of Variance (ANOVA) for the four water supply systems.

The previous overview shows that the management of drinking water resources implies the application of statistical methods and statistical software for general purposes but also application of specialized software solutions. Some of these solutions are listed below.

- Matlab/Octave toolbox for the application of GSA [7], called SAFE (Sensitivity Analysis For Everybody) is one of the Global Sensitivity Analysis(GSA) software tools from Matlab software package. Its increasingly used in the development and assessment of environmental models. SAFE is open source and freely available for academic and non-commercial purpose.
- Continuous water quality monitoring combined with web-based software [8] allows for an early warning of toxic algal blooms in lakes, seas, and desalination plants. A floating buoy system (Fig. 1) measures essential algae indicators (Chlorophyll-a, Phycocyanin, and Turbidity) and water quality parameters (Dissolved Oxygen (DO), Redox, pH, and Temperature) in order to monitor the water quality. The measured data can be viewed in real-time via a web-based software called the MPC-View.

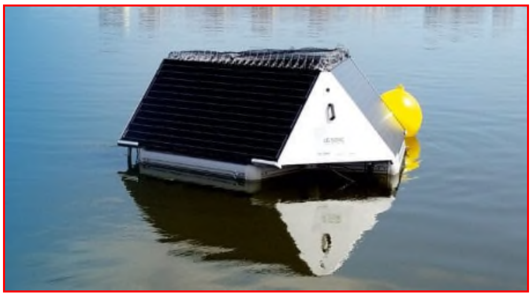
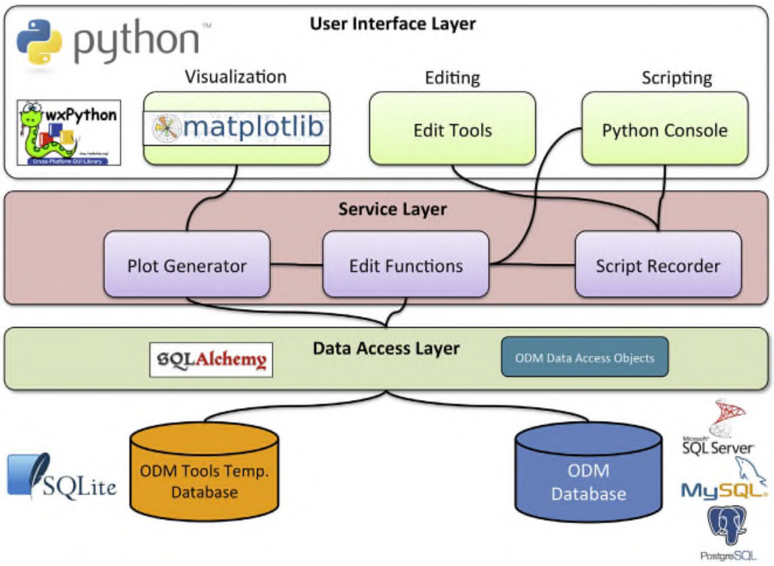


Figure 1: Real-time Water Quality Monitoring Software

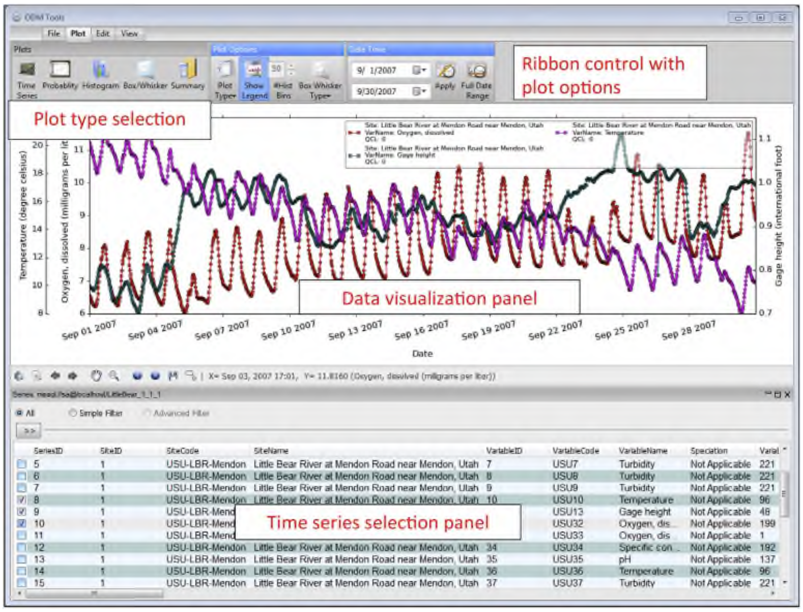
- WatPro™ [9] is the premier water treatment simulator for predicting water quality based on specific treatment processes and chemical addition (e.g. alum, ferric chloride, NaOH, lime). WatPro™ uses raw water quality parameters and operation parameters of process tanks, to simulate the plant performance.

In addition to the above examples, as well as commercial software solutions, such as MatLab and SPSS software packages, for optimal drinking water resources management open sourceplatforms are also very significant. As the most commonly used, we refer to Python and R.

- ODM Tools Python [10], is an open source software application that allows users to query and export, visualize, and perform quality control post processing on time series of environmental observations data stored in an ODM database using automated Python scripting that records the corrections and adjustments made to data series in the quality control process and ensures data editing steps are traceable and reproducible. The software architecture of ODM Tools Python is shown in Fig. 2. On Data storage layer, an ODM database implemented within an RDBMS that supports publication of observational data via standardized web services that query data from the ODM database in response to user requests and then return data in a standard XML schema called Water Markup Language (WaterML). ODM Tools Python uses a SQLAlchemy-based data access layer [11]. This layer serves to abstract data from the ODM database and provides a set of programmable objects that facilitate data management rather than repeatedly programming Structured Query Language (SQL) queries directly against to the ODM database. The service layer consists of a set of Python-based services containing the core functionality of the software application. The user interface layer provides the GUI within which users can visualize and export data, generate summary statistics, and perform data quality control editing. The GUI was designed and implemented using wxPython [12], which is a toolkit for Python that provides programmers with components for building interactive GUIs (Fig. 3).



- **Figure 2:** ODM Tools Python software architecture [10]



X= Sep 03, 2007 17:01, Y= 11.8160 (Oxygen, dissolved (milligrams per liter))

Series: mssql://sa@localhost/LittleBear_1_1_1

	SeriesID	SiteID	SiteCode	SiteName	VariableID	VariableCode	VariableName	Speciation	Varial
□	5	1	USU-LBR-Mendon	Little Bear River at Mendon Road near Mendon, Utah	7	USU7	Turbidity	Not Applicable	221
□	6	1	USU-LBR-Mendon	Little Bear River at Mendon Road near Mendon, Utah	8	USU8	Turbidity	Not Applicable	221
□	7	1	USU-LBR-Mendon	Little Bear River at Mendon Road near Mendon, Utah	9	USU9	Turbidity	Not Applicable	221
☑	8	1	USU-LBR-Mendon	Little Bear River at Mendon Road near Mendon, Utah	10	USU10	Temperature	Not Applicable	96
☑	9	1	USU-LBR-Mend			USU13	Gage height	Not Applicable	48
☑	10	1	USU-LBR-Mend			USU32	Oxygen, dis...	Not Applicable	199
□	11	1	USU-LBR-Mend			USU33	Oxygen, dis...	Not Applicable	1
□	12	1	USU-LBR-Mendon	Little Bear River at Mendon Road near Mendon, Utah	34	USU34	Specific con...	Not Applicable	192
□	13	1	USU-LBR-Mendon	Little Bear River at Mendon Road near Mendon, Utah	35	USU35	pH	Not Applicable	137
□	14	1	USU-LBR-Mendon	Little Bear River at Mendon Road near Mendon, Utah	36	USU36	Temperature	Not Applicable	96
□	15	1	USU-LBR-Mendon	Little Bear River at Mendon Road near Mendon, Utah	37	USU37	Turbidity	Not Applicable	221

- **Figure 3:** ODM Tools Python graphical user interface [12]

- The Rattle package (the R Analytical Tool To Learn Easily) [13] provides a graphical user interface specifically for data mining using R. It also provides a stepping stone toward using R as a programming language for data analysis in many domains and thus as a data miner in the domain of optimal and sustainable management of available water resources. Rattle specifically uses a simple tab-based concept for the user interface, capturing a work flow through the data mining process with a tab for each stage. This software can load data from various sources (CSV, TXT, ARFF, and ODBC connections to many data sources including MySQL, SQLite, Postgress, MS/Excel, MS/Access, SQL Server, Oracle, IBM DB2…). *Module for Exploratory data analysis* provides numerous numeric and graphic tools for exploring data. *Transform module* provides a number of the common options for transforming, including rescaling, skewness reduction, imputing missing values, turning numeric variables into categorical variables, and vice versa, dealing with outliers, and removing variables or observations with missing values. Rattle also provides a straight-forward interface to a collection of descriptive and predictive model builders available in R. The data miner draws heavily on methodologies, techniques and algorithms from statistics, machine learning, and data science (decision trees, boosting, random forests, support vector machines, generalized linear models, and neural networks). Rattle also provides collection of tools for evaluating and comparing the performance of models. This includes the error matrix (or confusion table), lift charts, ROC curves etc.

Based on the previously perceived importance of drinking water during the 21st century and insights into the state of the art in the domain of managament of drinking water (number of scientific publications, statistical and software support), the authors of this work have set themselves the goal to develop an application, using the C # programming environment, for statistical evaluation of the quality of raw water (*AQUA Statistic software*). Performance of one module of *AQUA Statistic* software is validated in the Case Study and data collected in the district of Zlatibor in the southwest part of the Republic of Serbia.

2. THEORETICAL BACKGROUND

The following section describes basic theoretical elements (one-way and two-way analysis of variance) based on which *AQUA Statistic software* module, was developed. In addition to the theoretical background, the property of raw water, which is the focus of this paper, is briefly described.

2.1. Analysis of Variance

In order to simultaneously investigate the equality of the arithmetic mean of several samples at once, a statistical method called an analysis of variance (ANOVA) is used. The point of the variance analysis is to explain the overall variability of the observed phenomenon to the constituent components (sources): the variance that is created under the influence of controlled factors, and so called residual variance, which occurs under the influence of others, uncontrolled factors [14].

2.1.1. One-way analysis of variance

An one-way analysis of variance explores the influence of one factor A with the r levels (treatments) on the variability of the observed phenomenon (variable X). The model of one-way analysis of variance is denoted by the equation

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{1}$$

where: X_{ij} - j-th observation, selected from the i-th set (sample), μ - the common mean of the observed samples, α_i - the effect of the i-th treatment, and ε_{ij}-random error. The model is valid if the assumptions are met: normality, homoscedasticity, random errors are on average equal to zero ($E(\varepsilon_{ij}) = 0$) and mutually independent, and the assumption of additivity is fulfilled. The Analysis of Variance examines the assumption H_0

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_i \ldots = \mu_r = \mu \tag{2}$$

in relation to the alternative hypothesis H_1: The arithmetic means of at least two sets differ from one another (Fig 5, Fig. 4).

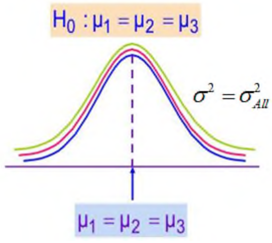


Figure 4. The zero hypothesis is accepted ($\sigma^2 = \sigma^2_{All}$)

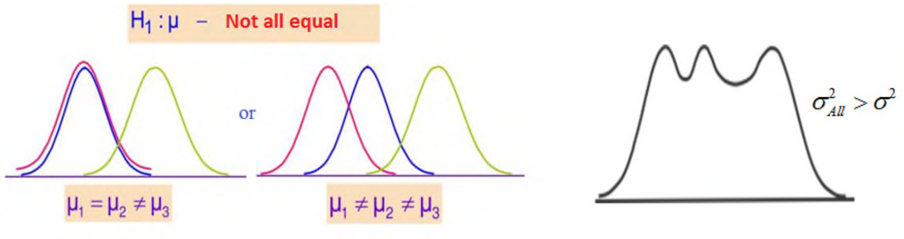

$H_1 : \mu$ – Not all equal
or
$\mu_1 = \mu_2 \neq \mu_3$
$\mu_1 \neq \mu_2 \neq \mu_3$
$\sigma^2_{All} > \sigma^2$

Figure 5. The zero hypothesis is rejected ($\sigma^2_{All} > \sigma^2$)
σ^2_{All} - variance of a common set, σ^2 - variance of individual sets

Analysis of Variance tests whether the variance between the groups is greater than the variance within the groups. If it is statistically significantly higher the zero hypothesis is not accepted, and vice versa. The ratio of variance between groups and variances within groups is tested by **F test** (Fisher test) (Eq.3) based on F statistics and Snedecor's F distribution (Fig. 6):

$$F = \frac{V_A}{V_R} \tag{3}$$

which represents the ratio between factor (V_A) and residual (V_R) variance:

$$V_A = \frac{S_A}{r-1} = \frac{n\cdot\sum_{i=1}^{r}(\overline{X}_i - \overline{\overline{X}})}{r-1},\ i=1,2,...,r \qquad V_R = \frac{S_R}{r\cdot n-1} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij}-\overline{X}_i)^2}{r\cdot n-1},\ i=1,2,...,r,\ j=1,2,...,n$$

where: n-the size of the sample, r-the number of samples, $r-1$- the number of degrees of freedom of factor variance, $r\cdot n-1$- the number of degrees of freedom of the residual variance, X_{ij}- j-th observation in the i-th sample, $\overline{X}_i$ - arithmetic mean of the i-th sample, and $\overline{\overline{X}}$ - is the common arithmetic mean of all the samples.
If the calculated **F** value is greater than the theoretical **F** value (**F$_{\text{critical}}$**), for the given level of statistical significance (α), it is concluded that the difference between the groups is statistically significant.

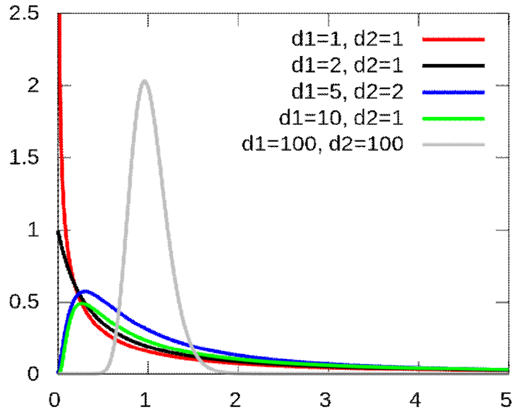


Figure 6. Snedecor's F distribution

In case where the realised (calculated) value of F statistics is in the critical area and when the samples have the same number of elements, the *Tukey* test is one of the most frequently used methods of multiple comparison that answers the question: Which sets have statistically significantly different arithmetic means? Tukey's test allows simultaneous comparison of all pairs of arithmetic mean of samples and is determined based on Tukey's critria,

$$T = Q_\alpha\sqrt{\frac{V_R}{n}} \tag{4}$$

where: Q_α - critical value of Tukey's test, V_R is a residual variance and n is the size of the sample. The calculated T criterion is compared with the absolute difference of the arithmetic mean of the samples [14]. If T is smaller, the arithmetic meanings differ significantly from one another, and if T is higher, the difference in the arithmetic mean of the samples is random, for the selected level of significance, α .

2.1.2. Two-way analysis of variance
When there are indications that the observed phenomenon is significantly influenced by several factors, analysis of variance models are applied with two or more factors. The model of the two-way analysis of variance is denoted by the equation:

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \tag{5}$$

where α_i, the effect i - th level of the factor A, β_j - the effect of the j-th level of the factor B, and the other members of the model are described in the text above.
In addition to the assumptions introduced for the one-way analysis of variance, for two-way analysis of variance, an additional assumption is introduced: Factors A and B are additive and there is no factor interaction. In addition, a two-way analysis of variance involves setting two different zero hypotheses: one by factor A, the other by factor B (Eq.6).

$$H_0:\alpha_1 = \alpha_2 = ... = \alpha_i ... = 0; \quad H_0:\beta_1 = \beta_2 = ... = \beta_i ... = 0 \tag{6}$$

A detailed procedure for a two-way analysis of variance is given in [14].

In practical research, the assumptions of variance analysis are very rarely met. However, in most scientific papers it is concluded that the deviation from normality, homogeneity and additivity will have little effect if the samples have the same size. In situations where samples (grops) significantly deviate from the normal distributio, it is recommended that the nonparametric alternative or Kruskal-Wallis test be applied.

2.2. Quality of raw drinking water

From a large number of raw water quality indicators, the validation of the developed software module, which is presented in this paper, was realized on the example of *electrical conductivity*, that is the ability of raw water to conduct an electric current (expressed in mS/m or $\mu S/cm$). Electrical conductivity depends on the concentration of ions in solution [15]. The dissolved solids are basically related to this measure, that is also influenced by the good conductivity of inorganic acids, bases, and the poor conductivity characteristic of organic compounds.

3. AQUA Statistic software modules

The employed *AQUA Statistic* software, which has been developed for statistical raw water quality assesment, was developed by the authors[1] in the Microsoft NET software environment (C# 6.0). The schematic view of *AQUA Statistic* module's structure (Unit 4) and its potential roles in the system of monitoring the quality of raw water in the Institute for Public Health, Uzice (PHU), are presented on Fig. 7.

From the appropriate water intakes (1), according to the prescribed standards, methods, and sampling dynamics, raw water samples are collected and analyzed in public health institutions (2). By means of measuring equipment, the measured values of chemical, physicochemical, physical, biological and other properties of raw water are stored in the database (3). MySQL DBMS was used to manage data. The *AQUA Statistic* software module consists of several units.

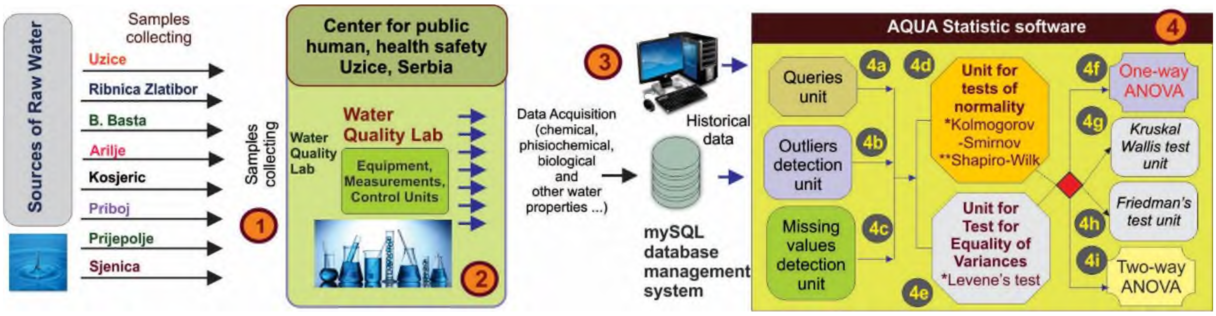

Sources of Raw Water
Samples collecting
Uzice
Ribnica Zlatibor
B. Basta
Arilje
Kosjeric
Priboj
Prijepolje
Sjenica
Samples collecting
1
Center for public human, health safety Uzice, Serbia
Water Quality Lab
Water Quality Lab
Equipment, Measurements, Control Units
2
Data Acquisition (chemical, phisiochemical, biological and other water properties ...)
3
Historical data
mySQL database management system
AQUA Statistic software
4
Queries unit
4a
Outliers detection unit
4b
Missing values detection unit
4c
4d
Unit for tests of normality *Kolmogorov -Smirnov **Shapiro-Wilk
Unit for Test for Equality of Variances *Levene's test
4e
4f
One-way ANOVA
4g
Kruskal Wallis test unit
4h
Friedman's test unit
4i
Two-way ANOVA

Figure 7: Schematic view of *AQUA Statistic* module's structure

Queries unit (4a) based on pre-defined queries or queries requested by the user will return the appropriate records. For data pre-processing, the *Outliers detection unit* (4b) and *Missing values detection unit* (4c) are employed. Component (4d, *Unit of Test of normality*) tests the normality of samples by means of Kolmogorov-Smirnov and Shapiro-Wilk's tests. Component (4e), *Unit for Test for Equality of Variances*) tests the homogeneity of variance by means of Levene's test. If the assumptions about the normality of the tested groups are fulfilled, the *One-way ANOVA* software unit (4f) is used for the analysis in two different regimes: a) *One-way between-groups ANOVA with post-hoc tests (Tukey HSD)* and b) *One-way ANOVA with repeated measures*. If the result of Levene's test had been significant (i.e., assumption not met), *One-way ANOVA* software uses an adjusted F test such as the Welch statistic and the Brown-Forsythe statistic. If the assumption of homogeneity of variance has been violated (Equal Variances Not Assumed), for the *post-hoc* comparison between groups, the Games-Howell or the Dunnett's C test, instead of Tukey HSD test, are used. If, however, the hypothesis of the normality of the tested groups is not met, a non-parametric alternative to the test can be used to compare the group (Kruskal Wallis test or Friedman's test, which is implemented through units (4g) and (4h)). Two-way ANOVA module (4i), carries out two-way, between-groups analysis of variance.

4. CASE STUDY

The following section presents the validation of the developed *AQUA Statistic* software module.

[1] *AQUA Statisitc* - authors: PhD Milovan Milivojevic, proffesor; Djordje Moljkovic and Miladin Tomic, students of third year on Informatic Tecnologies study program; Djordje Forst, software engineer specialist, assistant; Technical and Business College, Department for Informatics, Uzice, Serbia, 2017.

4.1. Location description

The historical data set was collected in the region of Zlatibor in the western part of Republic of Serbia, which together have over 320.000 inhabitants. Samples were collected from 8 water intakes (Fig. 8): Uzice (Source: Sušička vrela, Susica river), Cajetina (Reservoir on Ribnica River, Ribnica lake), Arilje (Rzav river), Kosjeric (Sources: Taorska vrela, Despotovica), Bajina Bašta (Source: by the river Drina), Priboj (Reservoir on the river Uvac - Radoinjsko lake).

4.2. Dataset. Variables and equipment

The historical data set ((3), Fig.7) was collected for the period of 01.01.2015. to 31.12.2016 (*dd.mm.yy* date formatting was applied). The database contained 814 records with values of attributes that define primary quality of raw water.
The *AQUA Statistic* software ((4), Fig. 2) has an input vector which contains 6 variables from which the quality of the primary characteristics of the raw water can be evaluated.

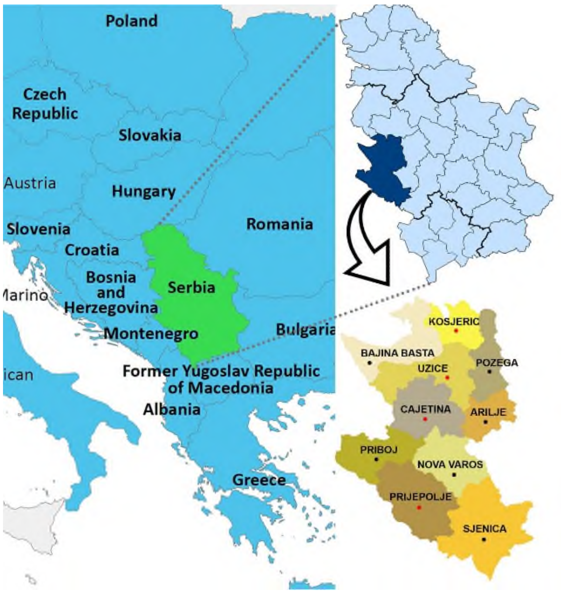


Figure 8: Locations of raw water intakes in the Zlatibor district, in the south-west part of Republic of Serbia.

The primary characteristics of raw water are indicated with the following variables: T, $Turbidity$, pH, Ec, $Oxid$, and $Chlor$, which represent: temperature, turbidity, pH value, electrical conductivity (20^0C), oxidizability - potassium permanganate consumption ($KMnO_4$) and chlorides (Cl^-), respectively. The historical data set was collected in *The Laboratory for quality control of water*, in the *Institute for Public Health, Uzice, Serbia* (Fig.9a, Fig. 9b). In this paper, for example, to assess the performance of *AQUA Statistic* software, the focus is on electrical conductivity. Measuring of raw water *electrical conductivity* were measured on the Hach USA Conductometer, type: Sension 7, ranges: 0-19.99μS/cm, 20-199.9μS/cm, 200-1999 μS/cm))/ method: EN 27888:1993. Indicators of the descriptive statistics of the treated data (Table 1) are shown, due to volume, only for sources for which the hypotheses on the normality of distribution have been confirmed (Kosjeric, Prijepolje, Cajetina-Ribnica, Uzice /water intake marked by red circles on Fig. 8).

Table 1. Historical dataset of raw water quality - descriptive statistics for electrical conductivity

	Kosjeric		Prijepolje		Ribnica		Uzice	
Mean	450.02		353.41		169.43		438.66	
Std.err. of Mean	4.70		2.88		3.09		3.81	
95% CI for Mean Lower Bound	443.34		348.31		162.98		432.82	
Upper Bound	462.24		359.80		175.35		447.90	
5% Trimmed Mean	452.04		353.95		168.96		440.75	
Median	445.00		350.50		171.00		444.00	
Variance	1036.04		597.26		583.01		1991.63	
Std. Deviation	32.19		24.44		24.15		44.63	
Minimum	401.00		300.00		123.00		309.00	
Maximum	522.00		403.00		217.00		555.00	
Range	121.00		103.00		94.00		246.00	
Interquartile Range	50.00		37.50		27.00		61.50	
Skewness	.383		.193		-.067		-.184	
Kurtosis	-.835		-.807		-.456		-.053	
	ID	Value	ID	Value	ID	Value	ID	Value
Max. values	1635	522.00	1627	403.00	151	217.00	4540	555.00
	1048	516.00	2357	399.00	4615	215.00	2503	532.00
	269	507.00	2523	398.00	4694	214.00	2413	527.00
Min. values	4147	401.00	763	300.00	2236	123.00	168	309.00
	3982	402.00	960	313.00	1375	128.00	826	342.00
	782	403.00	4190	316.00	839	129.00	443	346.00

Figure 9a: Institute for Public Health, Uzice, Serbia

Figure 9b: Quality of raw water process in Institute for Public Health, Uzice, Serbia,

4.3. Validation of AQUA Statistic software

This paper, in the text that follows, due to the limited space, primarily shows the AQUA Statistic module capabilities related to One-way and Two-Way follows analysis of variance.

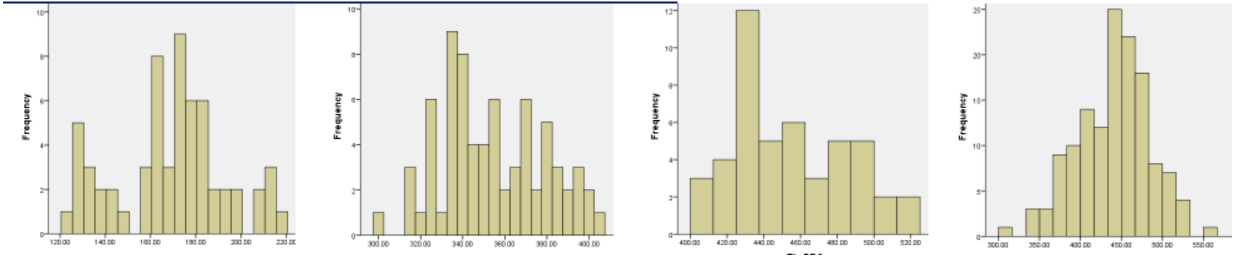
- *Testing the significance of electrical conductivity mean difference for the following water sources: Kosjeric, Prijepolje, Ribnica and Uzice with the use of the software module that carries out One-way between-groups ANOVA with post-hoc tests (Tukey HSD).*

In the first phase, assumptions of normality were tested on the complete dataset, using Kolmogorov-Smirnov and Shapiro-Wilk tests. Shapiro-Wilk test was chosen as a criterion to confirm the normality of distribution with a given significance level of $\alpha = 0.05$. Samples from four sources: Kosjeric (Sig. 0.073), Prijepolje (Sig. 0.109), Ribnica (Sig. 0.055), Uzice (Sig. 0.818) approximately follow a normal distribution (Table 2, Fig. 10).

Table 2. Test of Normality for the historical dataset of raw water quality: electrical conductivity

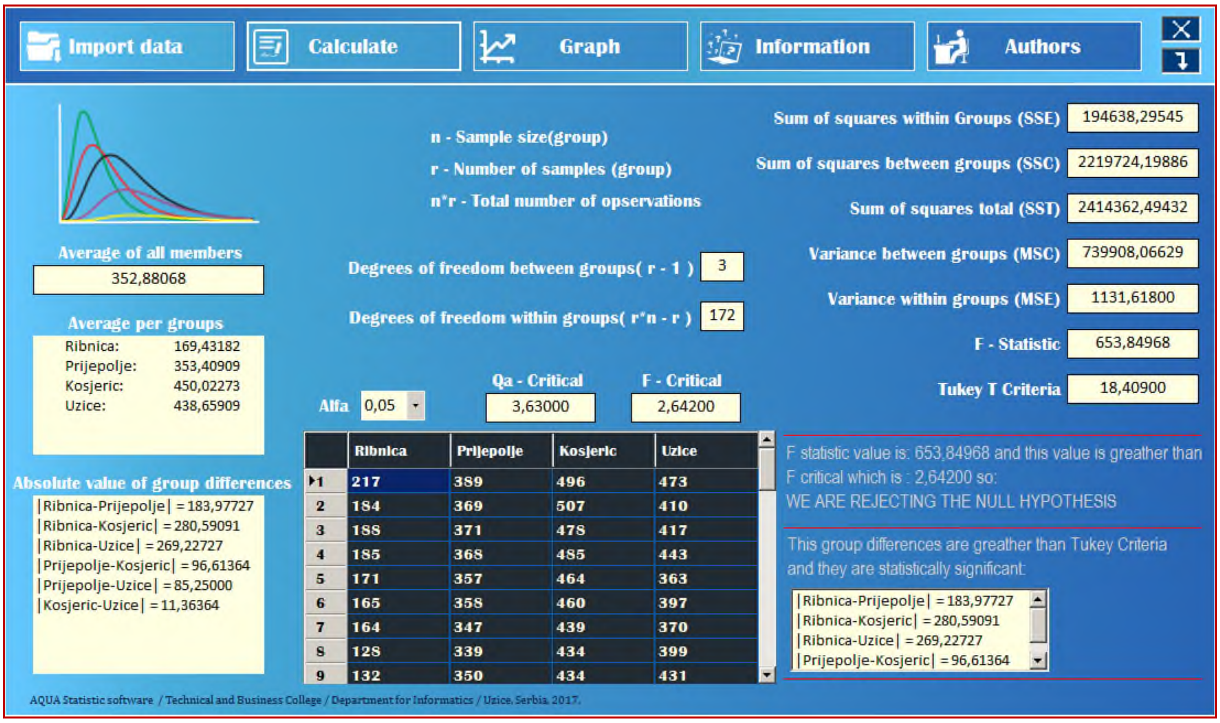
Water Intake	Kolmogorov-Smirnov		Shapiro-Wilk	
Location	*Statistic*	*Sig.*	*Statistic*	*Sig.*
Arilje	0.078	0.000	0.971	0.000
BBasta	0.148	0.002	0.958	0.038
Ribnica-Group1	0.089	0.200	0.962	**0.055**
Prijepolje-Group2	0.092	0.200	0.972	**0.109**
Priboj	0.103	0.069	0.958	0.020
Kosjeric-Group3	0.134	0.034	0.956	**0.073**
Sjenica	0.075	0.200	0.951	0.008
Uzice-Group4	0.067	0.200	0.994	**0.818**

Levene's test of Homogeneity of Variance did not confirm the assumption of equality of variances (Levene's statistic=8.53; df1=3; df2=172, Sig. 0.000). Adjusted F test (the Welch statistic - 857.168 and the Brown-Forsythe statistic - 653.850) values were calculated, but nevertheless the classical F statistic value was adopted - 653.850, because of equal sample sizes between groups (44 records), so the inequality of variance is of less importance. Module interface and test results are given in Fig. 11.



Ribnica *Prijepolje* *Kosjeric* *Uzice*

Figure 10. Distribution of Electrical conductivity [$\mu S / cm$] of raw water in Case Study



n - Sample size(group)
r - Number of samples (group)
n*r - Total number of opservations

Average of all members: 352,88068

Average per groups
Ribnica: 169,43182
Prijepolje: 353,40909
Kosjeric: 450,02273
Uzice: 438,65909

Degrees of freedom between groups(r - 1): 3
Degrees of freedom within groups(r*n - r): 172

Sum of squares within Groups (SSE): 194638,29545
Sum of squares between groups (SSC): 2219724,19886
Sum of squares total (SST): 2414362,49432
Variance between groups (MSC): 739908,06629
Variance within groups (MSE): 1131,61800
F - Statistic: 653,84968
Tukey T Criteria: 18,40900

Alfa: 0,05
Qa - Critical: 3,63000
F - Critical: 2,64200

Absolute value of group differences
|Ribnica-Prijepolje| = 183,97727
|Ribnica-Kosjeric| = 280,59091
|Ribnica-Uzice| = 269,22727
|Prijepolje-Kosjeric| = 96,61364
|Prijepolje-Uzice| = 85,25000
|Kosjeric-Uzice| = 11,36364

	Ribnica	Prijepolje	Kosjeric	Uzice
1	217	389	496	473
2	184	369	507	410
3	188	371	478	417
4	185	368	485	443
5	171	357	464	363
6	165	358	460	397
7	164	347	439	370
8	128	339	434	399
9	132	350	434	431

F statistic value is: 653,84968 and this value is greather than F critical which is : 2,64200 so:
WE ARE REJECTING THE NULL HYPOTHESIS

This group differences are greather than Tukey Criteria and they are statistically significant:
|Ribnica-Prijepolje| = 183,97727
|Ribnica-Kosjeric| = 280,59091
|Ribnica-Uzice| = 269,22727
|Prijepolje-Kosjeric| = 96,61364

AQUA Statistic software / Technical and Business College / Department for Informatics / Uzice, Serbia, 2017.

Fig. 11. AQUA Statistic - *One-way between-groups ANOVA with post-hoc tests (Tukey HSD) module*

A statistically significant difference between the electrical conductivity of water from these sources was determined at the level of $p < 0.05$: F(3,172)=653.849 is greater than $F_{critical} = 2.642$. The *effect size*, calculated using *eta squared*, was 0.919, which in Cohen’s [16] terms would be considered a large effect size. Cohen classifies .01 as a small effect, .06 as a medium effect and .14 as a large effect. Post-hoc comparisons using the Tukey HSD test indicated that the mean score for Group 1 (Ribnica) (M=169.43, SD = 24.15) was significantly different from Group 2 (Prijepolje) (M=353.41, SD=24.44), from Group 3 (Kosjeric) (M=450.02, SD=32.19) and from Group 4 (Uzice) (M=438.66, SD=44.63). Also, the electrical conductivity of raw water in Prijepolje is statistically significantly different from the electrical conductivity of the water in Kosjeric and Užice. Group 3 (Kosjeric) (M= 450.02, SD=32.19) did not differ significantly from Group 4 (Uzice) (M=438.66, SD=44.63). In other words, in terms of electrical conductivity, only the raw water in Uzice and Kosjeric are very similar, other tested raw drinking water sources from the Zlatibor region are very different.

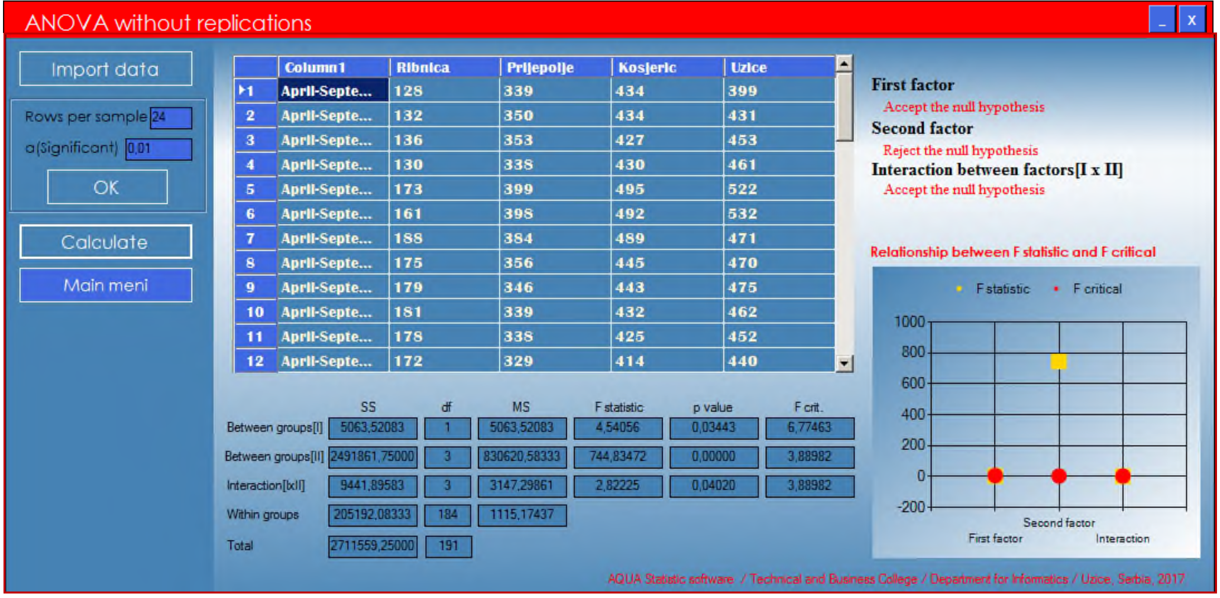
- *Testing the significance of electrical conductivity mean difference for the Uzice water source using the One-way repeated measures ANOVA module*

In order to show the performance of AQUA Statistic software in the domain of One-way repeated measures ANOVA, the significance of electrical conductivity mean difference by meteorological season was tested, for the period 2015-2016., for the water source Uzice. Each of the groups: Winter, Spring, Summer, Autumn, consisted of 36 records. Shapiro-Wilk test was chosen as a criterion to confirm the normality of distribution with a given significance level of $\alpha = 0.05$. Samples of all groups: Winter (M=427.56, SD=44.22, Sig. 0.130), Spring (M=434.64, SD=42.77, Sig. 0.248), Summer (M=447.92, SD=47.84, Sig. 0.523), Autumn (M=447.31, SD=40.62, Sig. 0.423) follow a normal distribution. By using the One-way repeated measures ANOVA it was determined that there wasn’t a significant effect seasonal effect on mean electrical conductivity values(Wilks’ Lambda = .826, F (3, 33) = 2.324, p =0.093>0.05).

- *Testing the significance of electrical conductivity mean difference by half-yearly periods (factor A), for the water sources: Ribnica, Prijepolje, Kosjeric, Uzice (factor B) with the use of Two-way ANOVA without repeated measurements module(Two-way between groups ANOVA)*

Two-way ANOVA without repeated measurements allows the examination of the basic effect of two independent variables, as well as their potential interactions, on the dependent variable. To showcase the performance of this AQUA Statistic software module (Fig. 12) an example was chosen, and the results are given in the text that follows.

A two-way between-groups analysis of variance was conducted to explore the impact of *Half-yearly Weather* (Weather) and *Raw Water Intake* (Intake) on Electrical conductivity for raw drinking water in the Zlatibor region. Time periods were divided into two levels: Dry (from April to September) and Wet (from October to March), for 2015-2016 interval. Raw water intakes were divided into four groups according to their location (Group 1: Ribnica; Group 2: Prijepolje; Group 3: Kosjeric; Group 4: Uzice).



ANOVA without replications

Rows per sample 24

a(Significant) 0,01

	Column1	Ribnica	Prijepolje	Kosjeric	Uzice
▸1	April-Septe...	128	339	434	399
2	April-Septe...	132	350	434	431
3	April-Septe...	136	353	427	453
4	April-Septe...	130	338	430	461
5	April-Septe...	173	399	495	522
6	April-Septe...	161	398	492	532
7	April-Septe...	188	384	489	471
8	April-Septe...	175	356	445	470
9	April-Septe...	179	346	443	475
10	April-Septe...	181	339	432	462
11	April-Septe...	178	338	425	452
12	April-Septe...	172	329	414	440

	SS	df	MS	F statistic	p value	F crit.
Between groups[I]	5063,52083	1	5063,52083	4,54056	0,03443	6,77463
Between groups[II]	2491861,75000	3	830620,58333	744,83472	0,00000	3,88982
Interaction[IxII]	9441,89583	3	3147,29861	2,82225	0,04020	3,88982
Within groups	205192,08333	184	1115,17437			
Total	2711559,25000	191				

First factor
Accept the null hypothesis
Second factor
Reject the null hypothesis
Interaction between factors[I x II]
Accept the null hypothesis

Relationship between F statistic and F critical

AQUA Statistic software / Technical and Business College / Department for Informatics / Uzice, Serbia, 2017

Fig. 12. AQUA Statistic - *Two-way between groups ANOVA module*

Levene’s Test of Equality of Error Variances provides a test of one of the assumptions underlying the analysis of variance. A significant result (Sig. value less than .05) suggests that the variance of variables across the groups is not equal. This holds true in our case study (Levene’s statistic=3.439; df1=7; df2=184, Sig. 0.002) and we set a more stringent significance level ($\alpha = 0.01$) for evaluating the results of two-way ANOVA, as recommended in [17].

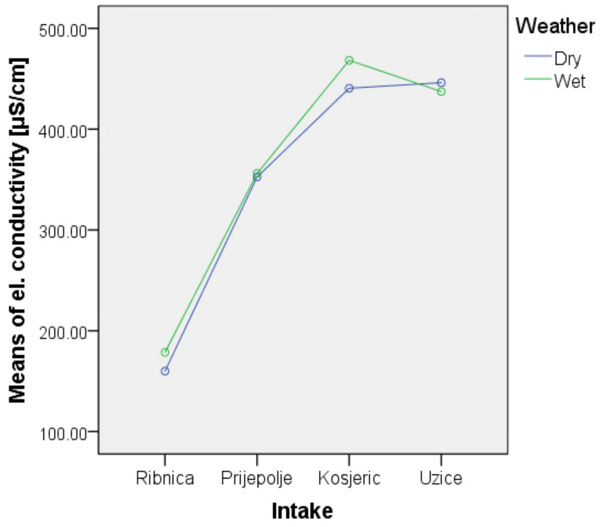


Fig. 13. *Two-way between groups ANOVA diagrams*

That is, we will consider the main effects and interaction effects significant only if the Sig. value is greater than .01. The interaction effect between Weather and Intake group was not statistically significant, F(3,184)=2.822, p=0.040. There was a statistically significant main effect for Intake, F(3,184)=744.83, p=0.000; and the effect size was large (partial eta squared =0.924). Post-hoc comparisons using the Tukey HSD test indicated the same results as *One-way between-groups ANOVA* with post-hoc test, which is described above. The main effect for Weather, F(1,184)=4.541, p=0.034, did not reach statistical significance.

In other words, Dry and Wet weather do not significantly influence the mean raw water electrical conductivity values, while there is a significant effect originating from the nature of the water source: Kosjeric and Užice have simular water, while the remaining water sources differ both from them, and each other. Plot in Fig. 13 is very useful for allowing users to visually inspect the relationship among variables.

5. CONCLUSION

Based on the importance of drinking water in the 21st century and insight into leading edge trends in the domain of management of drinking water (number of scientific publications, statistical and software support), this paper presents AQUA Statistic software for statistical evaluation of raw water quality, which the authors of this paper developed themselves, using the C# 6.0 programming environment. Performance of modules of AQUA Statistic software is validated in the Case Study and on the dataset of raw water properties, collected in the district of Zlatibor in the southwest part of the Republic of Serbia. Special emphasis is placed on software modules dealing with analysis of variance (One-way ANOVA, Two-way ANOVA). The performance of the developed software and the results of numerous accompanying statistical tests were validated on the example of raw water electrical conductivity. The developed software has some of the features of well-known commercial statistical packages. Advantages in relation to commercial packages are reflected in the possibility of automating the Integrated Raw Water Management and Monitoring System. In addition, it is possible to improve and integrate numerous Artificial Intelligence and Data Mining algorithms based on the open source platform (R, Python, …). However, commercial packages have a much larger set of tests, which is for AQUA Statistic software authors, the goal of future development.

Acknowledgement
The authors would like to thank the management, technologists, chemists, the engineers and technicians in the *Institute for Public Health, Uzice, Serbia* for their help and expert knowledge support in the monitoring technology and assessment of quality of raw drinking water in Zlatibor region. Also, the authors would like to thank them for collecting the data relevant for this study.

REFERENCES

[1] AMMAR, T.A.; ABID, K.Y.; EL-BINDARY, A.A.; EL-SONBATI, A.Z.: *Chlorine dioxide bulk decay prediction in desalinated drinking water*, Desalination, 352, 45–51, 2014.

[2] BEAUDEAU, P., SCHWARTZ, J., LEVIN, R.: *Drinking water quality and hospital admissions of elderly people for gastrointestinal illness in Eastern Massachusetts, 1998-2008*, Water Research, 52,188-198, 2014.

[3] SCHEILI, A., RODRIGUEZ, M.J., SADIQ, R.: *Seasonal and spatial variations of source and drinking water quality in small municipal systems of two Canadian regions*, Science of the Total Environment, 508, 514–524, 2015.

[4] DAHHOUA, M., MOUSSAOUITIA, M. E., MORHIT, M.E., GAMOUH, S., MOUSTAHSINE, S.: *Drinking water sludge of the Moroccan capital: Statistical analysisof its environmental aspects*, Journal of Taibah University for Science, 11, 749–758, 2017.

[5] ANTONOPOULOS, V.Z., PAPAMICHAIL, D.M., MITSIOU, K.A.: *Statistical and trend analysis of water quality and quantity data for the Strymon River in Greece*, Hydrology and Earth System Sciences, 5(4), 679–691, 2001.

[6] FABROA, A.Y.R., ÁVILA, J.G.P., ALBERICH, M.V.E., SANSORES, S.A.C., CAMARGO-VALERO, M.A.: *Spatial distribution of nitrate health risk associated with ground water use as drinking water in Merida, Mexico*, Applied Geography, 65, 49–57, 2015

[7] PIANOSI, F., SARRAZIN, F., WAGENER, T.: *A Matlab toolbox for Global Sensitivity Analysis*, Environmental Modelling & Software, 70, 80-85, 2015.

[8] https://www.lgsonic.com/water-quality-monitoring-software/

[9] https://www.linkedin.com/pulse/hydromantis-watpro-version-40-software-modeling-drinking-beres

[10] HORSBURGH J.S., REEDER S.L., JONES A.S., MELINE J., *Open source software for visualization and quality control of continuous hydrologic and water quality sensor data*, Environmental Modelling & Software, 70, 32-44, 2015.

[11] http://www.sqlalchemy.org

[12] http://www.wxpython.org

[13] https://journal.r-project.org/archive/2009-2/RJournal_2009-2_Williams.pdf

[14] ZIZIC, M., LOVRIC, M., PAVLICIC, D., Metodi statisticke analize, Ekonomski falutet, Beograd, 2006.

[15] DE ZUANE, J.: *Handbook of drinking water quality*, John Wiley & Sons, 1997.

[16] COHEN, J.W., Statistical power analysis for the behavioral sciences (2nd edition), Hillsdale, NJ: Lawrence Erlbaum Associates., 1988.

[17] PALLANT, J., SPSS Survival Manual (Third edition), Allen&Unwin, 2007.