DETERMINATION OF THE GEOMETRIC (*m*,*n*) STRUCTURE OF EXPERIMENTALLY PRODUCED CNTs

UDC: 514.123:546.24-022.532]: 004.92

Viktor Andonovic, Aleksandar T. Dimitrov, Perica Paunovic, Beti Andonovic

Abstract. Each carbon nanotube (CNT) has its own mathematical representation due to its hexagonal lattice structure. The subjects of research are multi-wall carbon nanotubes (MWCNTs) and determining their structural parameters: innermost and outermost diameters, chiral indices *m* and *n*, number of walls and their unit cell parameters. Within this paper low frequency region and corresponding high frequency parts of Raman spectra of three experimentally produced CNTs are considered, as well as Python programming for the most accurate (*m*,*n*) assignment. Determining the chirality of these samples enables calculation of other structural properties which are performed hereby. Furthermore, this author's work enables future studies on the samples, as are calculation of different topological indices using the graph representation and the chirality of the studied CNT samples.

1. INTRODUCTION

Carbon nanotubes (CNTs) are allotropes of carbon in nanodimensions with highly outstanding properties. Since graphene is a 2D building unit of all carbon allotropes, such as fullerenes, CNTs, nanoribbons, and so on, CNTs may be observed as wrapped up graphene structure having an ideal cylindrical shape, as shown in Fig. 1 A, B [4].

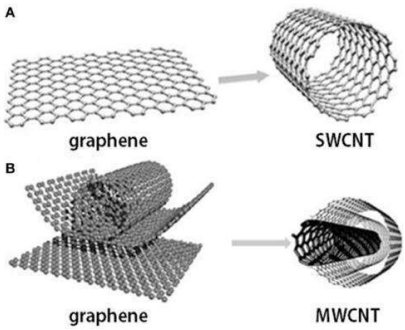

Figure 1: Graphene sheet as a 2D building unit of A) SWCNT; B) MWCNT.

2010 *Mathematics Subject Classification.* Primary: 51N20 Euclidean analytic geometry, 92E10 Molecular structure Secondary: 68-04 Computer science.
Key words and phrases. chiral indices, CNTs' structure, MWCNT, Python

Each nanotube has its own mathematical representation due to its hexagonal lattice structure [1],[2]. The geometric structure analysis of carbon nanotubes has been quite a challenging task, particularly if the subject of research is multi-wall carbon nanotubes (MWCNTs) [1]-[3], [5]-[8]. Knowing CNTs structural parameters (diameter, chiral angle, chiral indices m and n) is basically knowing their properties, which is essential for any research in the field of CNTs and their application. There are several excellent tools as are HRTEM, ED, RRS and others that suggest some models of (*m*,*n*) assignment for single-wall carbon nanotubes (SWCNTs), as well as for MWCNTs. However, precise determination of the CNTs atomic structure features becomes extremely complicated for more than three walls (layers) [5]. Thorough and overall analyses and use of experimental results combined with recent theoretical background may lead to successful estimation of its structural elements.

Within this paper Raman spectra of three experimentally produced CNTs of undetermined diameter, chirality, and number of walls, are considered, having assigned nomenclatures: CNT_1, CNT_2, and CNT_3. Due to the fact that their properties are tightly connected and dependent on their atomic structure, detailed analyses with regard to determining their diameters, calculating their chiral indices *m* and *n*, and furthermore other parameters, were calculated, hence estimating the number of walls (layers) of each nanotube. This research was strictly focused to determination of outermost and innermost diameters, as well as corresponding chiral indices, estimation of the number of other inner diameters, which would implicate the number and the nature of CNT's walls. Knowing the chirality of these samples enables calculation of other structural properties which are performed hereby. Authors strongly suggest future studies on the samples, as are performing additional EDP analysis to enhance and confirm the accuracy of applied methods, as well as calculation of different topological indices using the graph representation and the chirality of the studied samples, since it is known that they are related to some properties of the corresponding molecules.

2. MATERIALS AND TECHNIQUES

The carbonaceous phases extracted from the solidified electrolyte were observed by scanning electron microscopy, using JEOL 6340F (SEM, 10 kV). Structural characteristics of the carbon nanostructures were studied by means of Raman spectroscopy. Non-polarized Raman spectra were recorded by a confocal Raman spectrometer (Lab Ram ARAMIS, Horiba Jobin Yvon) operating with a laser excitation source emitting at 532 nm. The low frequency regions 50-350 cm^{-1} were taken into consideration, as well as the frequency regions 1200-1800 cm^{-1} from the Raman spectra of the CNTs. Python programming was applied to determine possible chiral indices assignment to the studied samples.

3. ANALYSES AND APPROACHES TO THE CNT SAMPLES

Knowing the mathematical representation of CNTs, due to its hexagonal lattice atomic structure, one can determine the relations among various CNT parameters: unit vectors $\vec{a}_1$ and $\vec{a}_2$, chiral vector $\vec{C}_h$, CNT diameter d_{CNT}, chiral angle θ, translation vector $\vec{T}$ of the CNT unit cell (being the shortest repeat distance along the nanotube axis), number of hexagons N_H, number of vertices (atoms) and so on, (Fig. 2), which are expressed with the formulas (1)-(5). Lattice constants are the lengths of the unit vectors $a = |\vec{a}_1| = |\vec{a}_2| = 0.246$ nm and the distances between neighbouring carbon atoms are $a_{C-C} = 0.142$ nm [1].

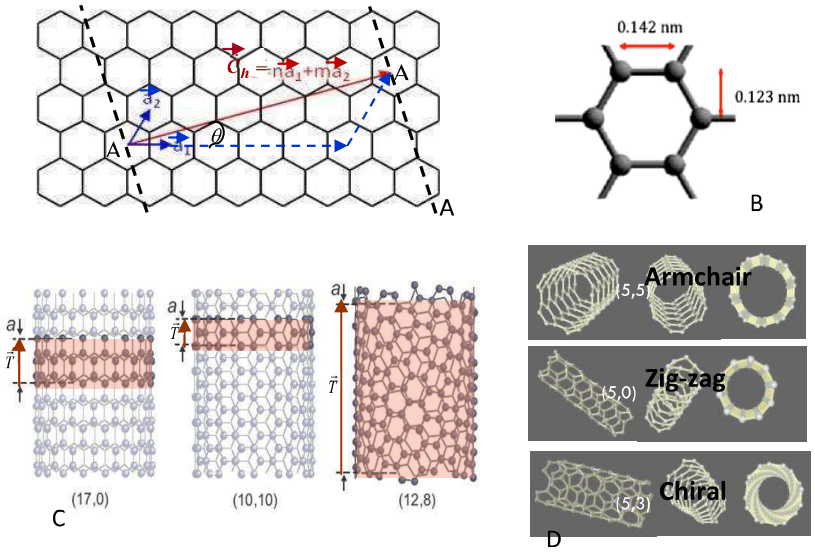

Figure 2: A) Atomic mathematical CNT structure; B) CNT constants; C) Various CNT unit cells; D)Three CNT types according to θ.

$$\left|\vec{C}_h\right| = \sqrt{3}a_{C-C}(m^2 + mn + n^2)^{1/2} \text{ nm} \tag{1}$$

$$d_{CNT} = \frac{\sqrt{3}a_{C-C}(m^2 + mn + n^2)^{1/2}}{\pi} = 0.079\sqrt{(m^2 + mn + n^2)} \text{ nm} \tag{2}$$

$$\theta = arctg\frac{\sqrt{3}n}{2m+n}, \quad 0 \le \theta \le \frac{\pi}{6} \quad (0 \le n \le m) \tag{3}$$

$$\theta = \frac{\pi}{6} \text{ – armchair CNT}, \quad \theta = 0 \text{ – zig-zag CNT}, \quad 0 < \theta < \frac{\pi}{6} \text{ – chiral CNT}$$

$$T = \left|\vec{T}\right| = \frac{\sqrt{3}\left|\vec{C}_h\right|}{d_R} = \frac{\sqrt{3}\pi d_{CNT}}{d_R} \text{ nm}, \tag{4}$$

$d_R = GCD(2m+n, 2n+m)$, given by

$$d_R = \begin{cases} d, & \text{if } m-n \text{ is not a multiple of } 3d \\ 3d, & \text{if } m-n \text{ is a multiple of } 3d \end{cases}$$

where $d = \text{GCD}(m,n)$

$$N_{\text{H}} = \frac{2(m^2+mn+n^2)}{d_R} \tag{5}$$

There are different experimental methods of producing CNTs, and based of the procedure, they can be either SWCNTs or MWCNTs. The CNT with the lowest reported diameter value experimentally produced is known to have the diameter $d_k = 4\text{Å}$ [10]. As theoretically predicted, it is the narrowest attainable that can still remain energetically stable. Such nanotubes may be the innermost constituent layer of MWCNTs. In contrast to CNTs with larger diameters, whose conductivity nature depends on their diameter and helicity (chirality), these smallest nanotubes are always metallic, regardless of the chirality [10]. For every other CNT it holds that it is metallic, if and only if it satisfies the condition $\text{MOD}(2m+n,3)=0$. In Fig. 3 images of experimentally obtained MWCNT samples at the Faculty of Technology and Metallurgy in Skopje are presented.

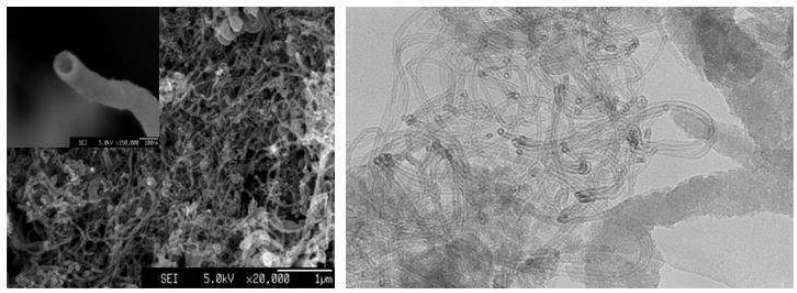

Figure 3: SEM and TEM images of experimentally obtained MWCNT

To date, the atomic geometric structure of carbon nanotubes determination and analysis has been quite a challenging task, particularly if the subject of research is MWCNTs. Three CNTs, experimentally produced, are considered for analyses in this research: CNT_1, CNT_2, and CNT_3. Each of the three nanotubes is undetermined with regard to its diameter, chirality and number of walls. The focus of this research is strictly focused to determination of their diameters, both

innermost and outermost diameters denoted by d_i and d_o correspondingly, determination of chiral indices (*m*,*n*), the total number and nature of other inner walls (layers), each having a diameter d_k, and the interlayer distance δ_r^k (Fig. 4). The latter are known to usually be within the interval $0.32\,\text{nm} \le \delta_r^k \le 0.35\,\text{nm}$, although it can sometimes vary from $0.27\,\text{nm}$ up to $0.35\,\text{nm}$ [9].

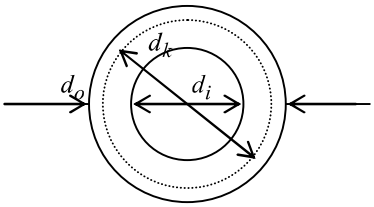

Figure 4: Diameters of MWCNTs

The relation among the diameters and the interlayer distances (Fig. 4) is given by

$$\delta_r^k = \frac{d_k - d_{k-1}}{2},$$

whereas by

$$N = \frac{d_o - d_i}{2\delta_r} + 1$$

the relation among the number of layers *N*, the innermost diameter d_i, the outermost diameter d_o, and the average interlayer distance δ_r is given.

The key role within our analyses was assigned to the Raman spectra in two frequency regions of each nanotube samples: $\omega \in [50\,\text{cm}^{-1}, 350\,\text{cm}^{-1}]$ and $\omega \in [1200\,\text{cm}^{-1}, 1800\,\text{cm}^{-1}]$. The first region was expected to point to the Radial breathing mode (RBM) frequencies at SWCNTs, and to Radial breathing-like mode frequencies (RBLM) at MWCNTs. These measurements can be used as an accurate tool to estimate the diameters of each layer of the tubes, since RBM is an active mode where all carbon atoms move in-phase in the radial direction. Several experimental relations have been established between the diameter of the tube and the RBM frequency ω_{RBM} [5]. Within a limited range of diameters, those relations are equivalent. While (6) is more accurate when small diameters are considered, (7) is more useful when it comes to very large diameters (or extremely low frequencies). The dependence of the innermost diameter d_i and the outermost diameter d_o on the corresponding frequency is given by the established relations in (6), whereas relation (7) can be used for much larger range of diameters, and whenever relation (6) is unusable or unreliable, due to the size of the outermost diameter. One may notice that d_i is obtained by the same equation as d_o, with C_e being 0. The latter is due to the fact that the parameter C_e is conventionally used to express different environmental conditions around the nanotube. The innermost concentric nanotube within the MWCNT is not affected by such conditions, which is not the case with the outermost concentric tube.

$$d_i = \frac{228}{\omega_i^{RBLM}}$$

$$d_o = \frac{228}{\sqrt{(\omega_o^{RBLM})^2 - 228^2 \cdot C_e}}, \qquad C_e = 0.065\ \text{nm}^{-2} \tag{6}$$

$$\omega_{RBM} = \frac{A}{d} + B \tag{7}$$

$A = 223\ \text{cm}^{-1}, \quad B = 10\ \text{cm}^{-1}$

Another frequency region to help the analysis is the high frequency region of G-modes, showing off in the interval $\omega \in (1500\ \text{cm}^{-1}, 1600\ \text{cm}^{-1})$. A useful diameter dependence for the chiral CNTs (semiconducting or metallic), which was used as a control method herein, is given by the formulas (8) and (9) [5].

$$\omega_{TO}^{G}(d) = 1582 - \frac{27.5}{d^2}, \quad \text{for semiconducting chiral} \tag{8}$$

$$\omega_{LO}^{G}(d) = 1582 - \frac{38.8}{d^2}, \quad \text{for metallic chiral} \tag{9}$$

The number of components in the G-peak is an excellent indicator of the conductive nature of the studied CNT sample [5]. Table 1 will help the analysis when the chiral indices assignment to the samples is done.

Table 1: Dependence between G-peak components and the CNT conductivity

Nanotube	Number of components	G-peak profile
Semiconducting chiral (SC)	2	LO, TO: narrow, symmetric
Metallic chiral (MC)	2	LO: broad, asymmetric TO: narrow, symmetric
Armchair	1	TO: narrow, symmetric
Semiconducting zigzag (SZ)	1	LO: narrow, symmetric
Metallic zigzag (MZ)	1	LO: broad, symmetric

3. RESULTS AND DISCUSSIONS

3.1. NANOTUBE CNT_1. (n,m) ASSIGNMENT RESULTS

The Raman spectra in both frequency regions for CNT_1 are shown in Fig. 5 A, B. There is only one peak in the low frequency region, which indicates that the nanotube is a single-wall, i.e. $d^{(1)} = d_o^{(1)}$. Using (6), it is obtained:

$d^{(1)} = d_o^{(1)} = 0.665\ \text{nm}$

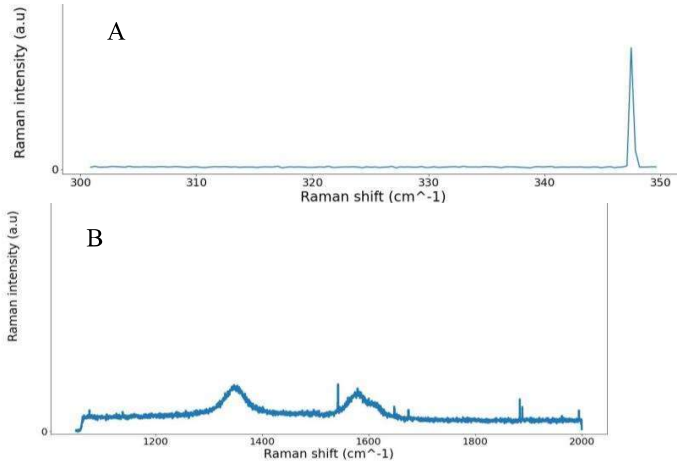

Figure 5: A) Raman spectrum RBM mode for CNT_1;
B) Raman spectrum G-mode for CNT_1

The determined diameter is lower than 1 nm and the most accurate results are expected for diameters 1 nm – 2.5 nm, hence the interval for calculating possible chiral indices candidates was performed within somewhat broader diameter interval $(d^{(1)} - 0.02, d^{(1)} + 0.02)$. Python programming was applied for obtaining possible candidates to satisfy equation (2), and the results of 24 possibilities are indicated in Table 2.

Table 2: Possible (*m*,*n*) assignments for diameter $d^{(1)} = 0.665$ nm of CNT_1

(4,3)	(4,4)	(5,2)	(5,3)	(5,4)	(5,5)
(6,0)	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)
(6,6)	(7,0)	(7,1)	(7,2)	(7,3)	(7,4)
(7,5)	(8,0)	(8,1)	(8,2)	(8,3)	(8,4)

Due to the broadness of the G-peak (see Fig. 5 B), the chiral indices candidate pairs need to satisfy the metallic condition $\mathrm{MOD}(2m+n,3)=0$. Hence, only the five chiral indices pairs in red (see Table 2) are considered. However, the armchair type (5,5) would show one narrow and symmetric G-peak, and the chiral metallic (6,3), (7,1), and (8,2) would show two components of the G-peak, one being narrow (see Table 1), which here is not the case.

This leaves only one possibility, the zig-zag metallic tube (6,0), which is in high accordance with the broad and asymmetric G-peak of the nanotube CNT_1.

In Figure 6 there is an illustration of CNT_1 with determined diameter and chiral indices assignment.

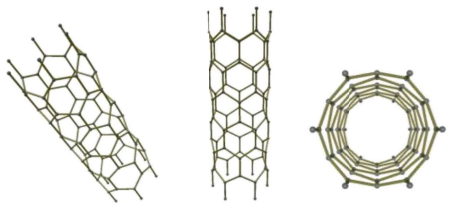

Figure 6: Visual model of (6,0) SWCNT

3.2. NANOTUBE CNT_2. (n,m) ASSIGNMENT RESULTS

The Raman spectra in PBLM frequency region, as well as in G-mode region for CNT_2 are shown in Fig. 7 A, B. There are two peaks in the low frequency range, which identifies the CNT_2 as double-wall CNT (DWCNT).

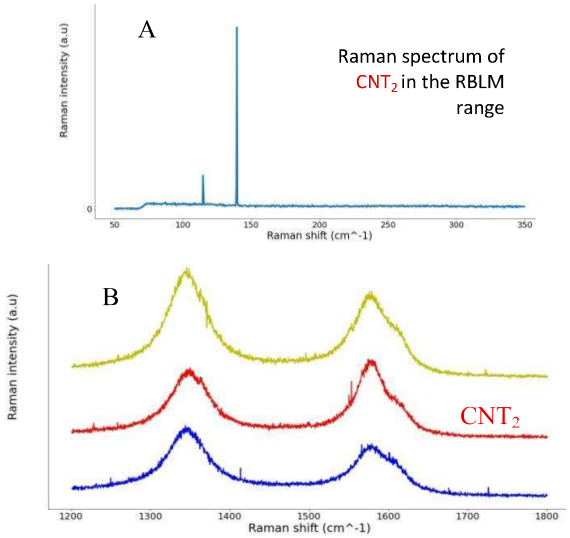

Figure 7: A) Raman spectrum RBLM modes for CNT_2;
B) Raman spectrum G-mode range for CNT_2

According to the corresponding frequency values in Fig.7 A and equation (6), both the innermost and outermost diameters are calculated as follows.

$d_i^{(2)} = 1.63$ nm

$d_o^{(2)} = 2.3\,\text{nm}$

The interlayer distance is calculated as $\delta_r = \frac{d_o - d_i}{2} = 0.335\,\text{nm}$, which is the equilibrium distance for tubes that are DWCNTs [5].

Considering the G-mode range, and depending on the chiral or achiral character of each constituent SWNT, one expects to observe 4 (chiral@chiral), 3 (chiral@achiral or achiral@chiral) or 2 (achiral@achiral) components in the Raman spectrum measured on an individual DWNT. However, some components can appear at close frequencies and thus cannot be experimentally resolved. Consequently, the number of observed components can be less than the one predicted or expected for different configurations (see Table 1). With regard to CNT_2, two identified components of the G-peak are located at the frequencies 1570.98 cm^{-1} and 1574 cm^{-1}. These frequencies enable the estimating of the diameters, and hence it is obtained $d_i^{(2)} = 1.58\,\text{nm}$ by using equation (8), which implicates a semiconducting chiral layer, and $d_o^{(2)} = 2.25\,\text{nm}$ by using equation (9), which implicates a metallic chiral layer. One may notice that the obtained diameters by both methods are in high accordance, and the values obtained by (6) are kept as accurate.

The determined diameters are within 1 nm – 2.5 nm, hence the interval for calculating possible chiral indices candidates was performed within narrower diameter interval with a 0.01 nm error bar $(d^{(2)} - 0.01, d^{(2)} + 0.01)$. Python programming was applied for obtaining possible assignment candidates for both diameters to satisfy equation (2), and the results of 16 combinations are derived from the pairs indicated in Table 3.

Table 3: Possible (*m*,*n*) assignments for the innermost and outermost diameters of CNT_2

$d_o^{(2)}$	$d_i^{(2)}$
(22,11)	(12,12)
(25,7)	(13,11)
(27,4)	(17,6)
(28,2)	(19,3)

Qualitative analysis of the G-peak indicated a broad component, hence a metallic chiral character of one layer and semiconducting chiral character of the other layer. Hence, the possible cases are either MC@SC or SC@MC, which corresponds to the implications from equations (8) and (9) results. The only chiral indices pair assignment satisfying the MC condition is (25,7). There are three possibilities of type SC@MC: (13,11)@(25,7), (17,6)@(25,7), and (19,3)@(25,7). The exact innermost and outermost diameters of these three possibilities give interlayer distances 0.329 nm, 0.335 nm, and 0,335 nm correspondingly. The latter eliminates the first tube, and, the remaining two possibilities (17,6)@(25,7), and (19,3)@(25,7) are equally possible up to here. However, it is possible to make a strong distinction between these two

possibilities by an additional method of performing and analyzing an EDP of the CNT_2. Authors suggest such further research, since this analysis would estimate the ratio m/n, which greatly differs at the last two possibilities. It is highly expected that it would leave only one candidate combination.

In Fig. 8 there is an illustration of the two candidates with determined diameters and chiral indices assignment to the tube CNT_2.

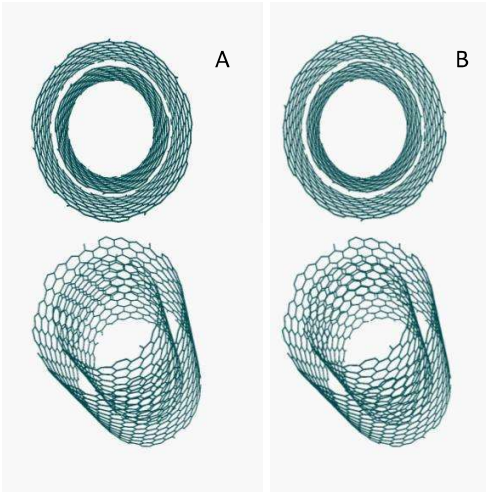

Figure 8: Visual model of A) (17,6)@(25,7); B) (19,3)@(25,7)

3.3. NANOTUBE CNT_3. (n,m) ASSIGNMENT RESULTS

Several (six) peaks may be noticed in the Raman spectra in PBLM frequency region of CNT_3 (Fig. 9 A), which identifies this nanotube as MWCNT. Presence of both narrow and broad components in the G-mode range (Fig. 9 B) implicates both semiconducting and metallic layer in the structure of CNT_3. According to the corresponding frequency values in Fig.9 A and equation (6), both the innermost and outermost diameters are calculated as follows:

$$d_i^{(3)} = 0.65 \text{ nm}$$

$$d_o^{(3)} = 7.73 \text{ nm}$$

The frequency of the outermost diameter $d_o^{(3)}$ is extremely low and near the limit of possible calculation, therefore its calculation may have a high error bar or even be highly inaccurate. Hence, the outermost diameter is recalculated using equation (7) and following result is obtained:

$$d_o^{(3)} = 4.04 \text{ nm}$$

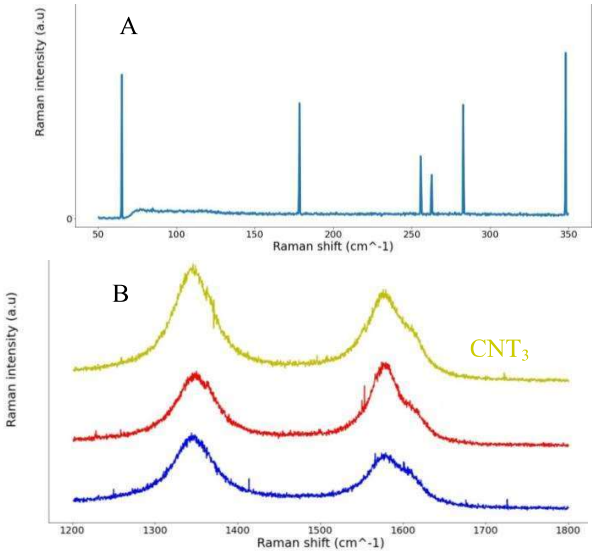

Figure 9: A) Raman spectrum RBLM modes for CNT_3;
B) Raman spectrum G-mode range for CNT_3

To be able to decide which outermost diameter value is good, equation (8) is applied for the G-peak component frequency $\omega = 1580.34\,\mathrm{cm}^{-1}$. Thus it is obtained $d_o^{(3)} = 4.07\,\mathrm{nm}$. The latter is in accordance with the calculated value by (7). Furthermore, the use of this equation points to semiconducting chiral layer. It is possible to estimate whether the choice of these equations was justified by checking the accordance with $N = \dfrac{d_o - d_i}{2\delta_r} + 1$, when calculating the average interlayer distance δ_r in CNT_3. Using the values N=6, $d_o^{(3)} = 4.04\,\mathrm{nm}$, and $d_i^{(3)} = 0.65\,\mathrm{nm}$, thus obtaining $\delta_r = 0.339\,\mathrm{nm}$, is an excellent indicator that the diameters are well estimated. The innermost and the outermost diameters are out of the high accuracy diameter range 1-2.5 nm, and obtained by different equations. The possible (*m*,*n*) assignment was performed using Python programming, equation (2) and corresponding error bars.

For the outermost diameter, there were three possibilities in the interval $(d_o^{(3)} - 0.002, d_o^{(3)} + 0.002)$: (31,28), (32,27), and (49,4), each being chiral. However, considering the semiconducting nature of this layer, the only pair of chiral indices satisfying the condition $\mathrm{MOD}(2m+n,3) \neq 0$ is (32,27).

With regard to the innermost diameter, the possible (*m*,*n*) assignment was performed in the interval $(d_i^{(3)} - 0.02, d_i^{(3)} + 0.02)$ resulting into two possibilities: (7,2) and (8,0), of which it must be emphasized that the pair (7,2) is within a much lower error bar. Both of these possibilities point to a semiconducting layer.

There are two possible combinations: (7,2)@@(32,27) and (8,0)@@(32,27). The option (7,2)@@(32,27) was discussed to be within a smaller error bar, and also its calculated interlayer distance of 0.340 nm is closer to the determined $\delta_r = 0.339\,\text{nm}$, than the distance of 0.341 nm that holds for the option (8,0)@@(32,27). However, these findings are not enough of a discrepancy at the latter option in order to be excluded.

In Fig. 10 there is an illustration of the two candidates with determined diameters and chiral indices assignment to the tube CNT_3.

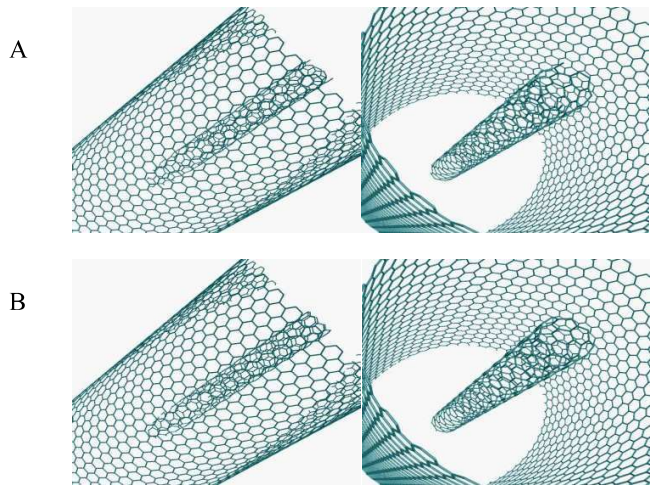

Figure 10: Visual model of A) (7,2)@@(32,27); B) (8,0)@@(32,27)

Considering this nanotube CNT_3 it is again strongly suggested an additional EDP analysis to be performed. This would differentiate the two candidates, since they have different *m*/*n* ratios of the innermost constituent tubes, and hence it is highly expected that the EDP would leave only one possible candidate.

4. CONCLUSIONS

Based on the analysis aned discussions in previous sections, several important findings may be concluded:

- Determination of three CNTs atomic structure has been performed; fully for CNT_1 and CNT_2, and partially for CNT_3 (as presented in Table 4);
- The thorough analyses were made with regard to the CNTs' Raman spectra in RBLM and G-mode frequency regions, combined with use of Python programming;
- The performed calculations were in excellent agreement with the theoretical background and with control methods;
- The calculations can be further improved in terms of higher accuracy, and corresponding methods, as EDP, are strongly suggested;
- The results enable many applications, as well as providing full necessary information for graph theorists who work on topological indices.

Table 4: Summarized results to specify studied CNTs

Parameters in nanotube	**CNT_1**	**CNT_2**	**CNT_3**
Chiral indices (*m*,*n*)	(6,0)	(17,6)@(25,7) (19,3)@(25,7)	(7,2)@@(32,27) (8,0)@@(32,27)
Diameters (in nm)	0.665	1.63@2.3	0.65@4.04
Interlayer distances (in nm)	-	0.335	0.34
Number of walls	1	2	6
Conducting nature	MZ	SC@MC	SC@@SC (intrinsically Metallic)
Chiral angles (in rad)	0	0.25@0.21 0.13@0.21	0.21@@0.47 0@@0.47
Length of the unit cell (in nm)	0.6	8.87@4.17 8.87@4.17	3.54@@21.98 0.44@@21.98
Number of hexagons in the UC	12	854@566 854@566	134@5234 16@5234

COMPETING INTERESTS

Authors have declared that no competing interests exist.

ACKNOWLEDGMENTS

This study was done within the following EU projects: COST Action CA17139 "European Topology Interdisciplinary Action", and COST Action CA17140 "Cancer nanomedicine - from the bench to the bedside".

References

[1] M.S. Dresselhaus, G. Dresselhaus, R. Saito, A. Jorio, *Raman spectroscopy of carbon nanotubes*, Physics Reports 409 (2005), 47–99.
[2] X. Zhao, Y. Ando, L.-C. Qin, H. Kataura, Y. Maniwa, R. Saito, *Radial breathing modes of multiwalled carbon nanotubes*, Chem. Phys. Lett. 361 (2002), 169–174.
[3] C. Schwandt, A.T. Dimitrov, D.J. Fray, *High-yield synthesis of multi-walled carbon nanotubes from graphite by molten salt electrolysis*, Carbon 50 (2012), 1311–1315.
[4] B. Andonovic, A. Ademi, A. Grozdanov, P. Paunović, Aleksandar T. Dimitrov, *Enhanced model for determining the number of graphene layers and their distribution from X-ray diffraction data*, Beilstein J. Nanotechnol 6 (2015), 2113-2122.
[5] D.I. Levshov, H.N. Tran, M. Paillet, R. Arenal, X.T. Than, A.A. Zahab, Y.I. Yuzyuk, J.-L. Sauvajol, T. Michel, *Accurate determination of the chiral indices of individual carbon nanotubes by combining electron diffraction and Resonant Raman spectroscopy*, Carbon 114 (2017), 141-159.
[6] T. Natsuki, G.J.H. Melvin & Q.-Q. Ni. *Vibrational Frequencies and Raman Radial Breathing Modes of Multi-Walled Carbon Nanotubes Based on Continuum Mechanics*, Journal of Materials Science Research 2, (4) (2013).
[7] J.M. Benoit, J.P. Buisson, O. Chauvet, C. Godon, and S. Lefrant. *Low-frequency Raman studies of multiwalled carbon nanotubes: Experiments and theory*, Phys. Rev. B, 66 (7) (2002), 073417.
[8] B. Andonovic, V. Andova, T. Atanasova Pacemska, P. Paunovic, V. Andonovic, J. Djordjevic, & A. Dimitrov, *Distance based topological indices on multiwall carbon nanotubes samples obtained by electrolysis in molten salts*. BJAMI, 3(1), (2020), 7-12.
[9] O.V. Kharissova, B.I. Kharisov, *Variations of interlayer spacing in carbon nanotubes*, RSC Adv., 4, (2014), 30807-30815
[10] L. Qin, X. Zhao, K. Hirahara, et al, The smallest carbon nanotube, Nature 408, 50, (2000), 50-51.

Jožef Stefan Institute, Ljubljana, Slovenia
E-mail address: viktor.andonovikj@ijs.si
Faculty of Technology and Metallurgy, University "St Cyril and Methodius", Skopje, Macedonia
E-mail address: aco@tmf.ukim.edu.mk
Faculty of Technology and Metallurgy, University "St Cyril and Methodius", Skopje, Macedonia
E-mail address: pericap@tmf.ukim.edu.mk
Faculty of Technology and Metallurgy, University "St Cyril and Methodius", Skopje, Macedonia
E-mail address: beti@tmf.ukim.edu.mk

REGRESSION METHODS IN ANALYSIS OF GEOTECHNICAL PARAMETERS OF COAL DEPOSITS

UDC: 519.242:553.94.048

Bojana Nedelkovska, Igor Peshevski, Milorad Jovanovski,
Daniel Velinov, Zoran Misajleski

Abstract. In this paper, a comparative analysis of geotechnical parameters of coal deposits from different sedimentation basins in Macedonia is given. The knowledge of the geotechnical properties of the rock masses in mineral deposits is very important for the economy of the extraction process, safety of work and protection of the environment. From the geological and geotechnical investigations of the coal deposits, usually a relatively large stock of data for the physical and mechanical properties of materials is obtained, forming a series that can be statistically analyzed. Analysis of selected samples and types of performed tests for several deposits was conducted, as well as a comparative analysis of the values for the following geotechnical parameters: natural moisture on each deposit, unit weight and coefficients of filtration. Several correlation dependencies have been established between certain geotechnical parameters, such as: unit weight - dry unit weight and porosity - filtration coefficient, for which high values for the correlation coefficient R^2 (which indicates a strong to very strong dependance) are obtained. With such analyses, a procedure for processing a larger pool of data from geotechnical investigations of mineral deposits is defined. The introduction of such a procedure aims in selecting and adopting the most relevant value for certain geotechnical parameter, and generally define reliable geotechnical parameters for coal deposits, which in turn are needed to perform a number of other geotechnical (both geological and mining) analyses.

1. INTRODUCTION

This paper is devoted on coal deposits and establishing a procedure for defining their most realistic geotechnical parameters. The importance of complete information for the geotechnical parameters is seen both in economic terms and safety of the extraction process, as well as environmental protection. Whether the coals are excavated underground or on the surface, it is always necessary to have detailed knowledge on their geotechnical parameters in order to have good implementation of the processes of planning, designing, exploitation and closing of coal mines. The coal is one of the main source of energy in our country, and in the last four decades some serious activities in coal's investigations, including geotechnical activities, are conducted.

2010 *Mathematics Subject Classification.* Primary: 62P30, Secondary: 65C20, 65C60.
*Key words and phrases.*coal deposit, geotechnical parameters, statistical analysis, regression, comparative, correlations.

The data of the state's electricity production company show that there are total confirmed reserves of 664 million tons of coal. This fact implies that there is strong need of formation of clear and detailed geotechnical profiles of the deposits, with emphasis on possible excavation depths in underground excavation or high slopes in the case of surface excavation. Therefore, the need of knowledge about the differences in geotechnical parameters in mineral deposits from different sedimentation basins and specific correlations used in different phases of design and exploitation of the coal, is quite strong.

Herein is presented analyzed data from four coal deposits. More precisely, data on natural moisture content, unit weights, and coefficient of filtration is considered. Then a correlation analysis between unit weight-dry unit weight and porosity-coefficient of filtration for certain materials within the coal deposits is conducted. As main and ultimate goal of this paper is to use regression method for proper selection and adoption of most representative values of specific geotechnical parameters, which can be used for further geotechnical analysis. Another goal in this paper is to perceive the influencing factors on the parameters, to compare and make possible correlations.

From all coal reserves, it is estimated that around 38% can be exploited with surface excavation, and the rest with underground technology (underground coal excavation in our country is still not applied). Having on mind that the surface excavation is the usual way of excavation in our country, it is necessary to have surface mines with relatively great depths. This includes design of stabile slopes, which on the other hand implies need for detailed picture for the geological and geotechnical profile of the coal deposit. Hence, a complex geological, engineering-geological, hydro-geological and geotechnical investigations, consisting of drilling exploratory boreholes, field and laboratory tests, have to be done.

For the purpose of the comparative analysis of the geotechnical parameters of coal deposits of different sedimentation basins and available data, a choice of certain mineral deposits, who will be considered in this paper is made. The chosen coal deposits are:

- Suvodol and Brod-Gneotino (part of the Pelagonia basin)
- Lavci (part of Prespa basin)
- Zvegor-Stamer (part of Delchevo-Pehchevo-Berovo's basin).

For all mentioned coal deposits, in different time periods, certain amount of geological and geotechnical, as well as laboratory works are made.

The data in this paper is taken from technical documentation prepared in the time period 1985 to 2018, and the number of the results for each coal deposit varies from 600 to 1200. We investigate only three parameters of the deposit materials in this paper. The large amount of data allows us to make statistical analysis with quite big confidentiality and also statistical comparison between different coal deposits. The comparative analysis is given by tables and graphs, usual tool of descriptive statistics, relieving the process of making conclusions. This type of approach allows to determinate the confidence interval for

particular geotechnical parameter, to detect non logical values, to find maximal, average and minimal values, standard deviations, median and other statistical parameters for the materials in the mineral deposits. Using these type of analysis, we can deduce certain conclusions for geotechnical parameters of particular deposit/s, to plan future phases of investigations for each deposit and to prescribe the number of sufficient number of tests. Correlations between certain parameters and dependences with high coefficient of correlation are obtained.

These type of correlations enable efficient programmed investigations for further explorations and reducing of unnecessary costs in case of consistency of some specific parameter. Having all this on mind, it is quite clear that the results placed here could be used in coal exploration and exploitation practices.

2. METHODOLOGY

In this section are presented the main statistical definitions and methods used for analysis of the data obtained from coal deposits mentioned in the Introduction. It important to note that, to our knowledge, this kind of approach in analysis of geotechnical parameters of the coal deposits has never been done before.

The definitions and basic facts about average values, standard error, median, mode, standard deviation, variance are quite known statistical and very recently used notations in descriptive statistics, so we are going to omit them. We are going to present same basic facts about skewness, kurtosis, confidence intervals, boxplots, correlation and regression.

The skewness usually is described as a measure of the symmetry of data, or it can be interpreted as a lack of symmetry. Perfectly symmetric set of data have skewness 0. The formula calculating skewness is the following

$$a_3 = \sum \frac{(x_i - \bar{x})^3}{ns^3}.$$

Here we can notice that the exponent is 3, since the skewness is "the third standardized central moment in the probability".

The kurtosis is statistical measure defining how the tails (the parts which are far from the peaks, i.e. the left and the right ends of the curve) of the distribution differ from the tail of standard normal distribution. This means that the kurtosis gives us information whether the tails of certain distribution contains extreme values. The kurtosis does not study the peaks (extremes), curved and flat parts of the distribution. The kurtosis is given by the following

$$a_4 = \sum \frac{(x_i - \bar{x})^4}{ns^4}.$$

It can be noticed that the exponent is 4, since the kurtosis is "the fourth standardized central moment in the probability".

In the expressions for skewness and kurtosis, n is the size of data, x_i is the i-th value of the sample, $\bar{x}$ is the average value of the sample data and s is the standard deviation.

Usually confidence intervals are used when we want to give an estimate of some parameter of the population, i.e. how accurate is the estimate with certain probability. 95% confidence interval means that when we have repeatedly sampling (sample series) from the population, 95% of the obtained intervals will include the real value of the parameter of the investigated population. It is clear that bigger sample implies more reliable confidence interval.

The maximal and minimal observations tell us little about the distribution of the sample values, but they give us information about the tails of the distribution that are missing if we know only the median and the quartiles. To obtain a quick summary of both center and spread, we combine the following five numbers: minimum, first quartile, median, third quartile and maximum, written in order from smallest to largest. These five numbers offer a reasonably complete description of center and spread. A boxplot is a graph of the five-number summary: A central box spans the quartiles Q_1 and Q_3, a line in the box marks the median M, lines extend from the box out to the smallest and largest observations. Because boxplots show less detail than histograms or stemplots, they are best used for side-by-side comparison of more than one distribution.

The regression line summarizes the relationship between two variables, but only in a specific setting when one of the variables helps explain or predict the other. A regression line is a line that describes how a response variable y changes as an explanatory variable x changes. This line often is used to predict the value of y for a given value of x. The least squares regression line of y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

A quadratic regression is the process of finding the equation of the quadratic function (parabola) that best fits a set of samples. As a result, we get an equation of the form

$$y = ax^2 + bx + c, \quad a \neq 0.$$

The best way to find this equation is by using the least squares method. That is, we need to find the values a, b and c such that squared vertical distances between each point (x_i, y_i) and the quadratic curve $y = ax^2 + bx + c$ is minimal. The equation of the curve of the regression curve can be found as a solution of the following system

$$\begin{cases} a \cdot \sum x_i^4 + b \cdot \sum x_i^3 + c \cdot \sum x_i^2 = \sum x_i^2 y_i \\ a \cdot \sum x_i^3 + b \cdot \sum x_i^2 + c \cdot \sum x_i = \sum x_i y_i \\ a \cdot \sum x_i^2 + b \cdot \sum x_i + n \cdot c = \sum y_i \end{cases}, \; i = 1, 2, ...n.$$

The coefficient of correlation R^2, can be found by the following formula

$$R^2 = 1 - \frac{RSS}{TSS},$$

whereby RSS is denoted the sum of squares of residuals and by TSS is denoted the total sum of squares, i.e.

$$RSS = \sum\left(y_i - ax_i^2 - bx_i - c\right)^2 \text{ and } TSS = \sum\left(y_i - \bar{y}\right)^2.$$

Correlation and regression are closely connected. It is important to note that we should be careful when we want to conclude that there is a cause-and-effect relationship between two variables just because they are strongly associated, i.e. high correlation does not imply causation.

According to the quantity of the coefficient of correlation (R^2) we can determine the strength of the regression, usually using the following criteria:

- if $R^2 < 0,3$ there isn't any dependence;
- if $0,3 \le R^2 < 0,5$ there is some dependence;
- if $0,5 \le R^2 < 0,7$ there is mild dependence;
- if $0,7 \le R^2 < 0,9$ there is strong dependence;
- if $R^2 \ge 0,9$ there is very strong dependence.

3. APPLICATIONS

In this part, comparative analysis for certain geotechnical parameters (natural moisture content, unit weight and coefficient of filtration) for certain type of geotechnical materials was made. There are given boxplot and certain tables for the geotechnical parameters for certain materials (average value, standard error, median, mode, standard deviation, variance, kurtosis, skewness, range, minimum, maximum, sum, sample numbers and confidence interval) and certain correlations between unit weight-dry weight of certain materials and correlations between porosity-coefficient of filtration for certain materials in coal deposits.

3.1. Natural moisture content

On the given boxplot above for each coal deposit are given minimal and maximal value of natural moisture, first and third quartile and median. For the coal deposits Suvodol, Lavci and Brod-Gneotino, minimal value of the moisture of the samples varies around 10% and maximal value varies around 50%. These values for the samples are quite different in the case of coal deposit Zvegor-Stamer, i.e. minimal value of the natural moisture around 5% and maximal value of the natural moisture content around 70%. Generally, we can

note that in all investigated coal deposits, the samples have natural moisture content between $20-40\%$.

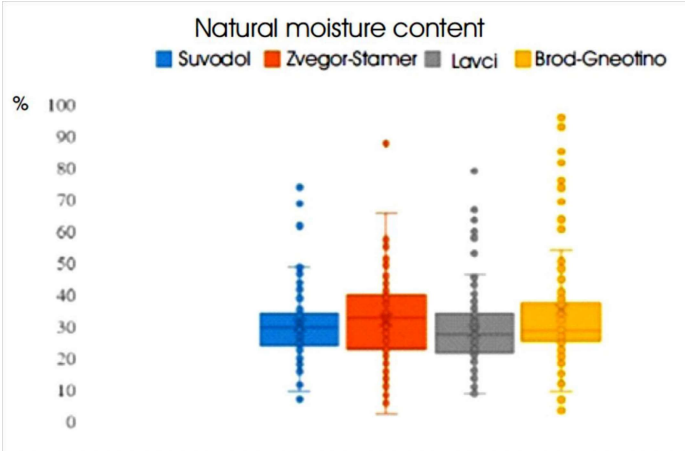

Figure 1. Natural moisture content of some deposits

3.2. Unit weight

Unit weight is one of the most often investigated geotechnical parameters in all deposits. This parameter is essential for every geotechnical analysis. The comparative analysis for all investigated coal deposits is given through the tables in sequel.

Table 1. Unit weight of gravels (in kN/m^3)

Gravels	**Zvegor-Stamer**	**Lavci**
Average value	19,20	20,48
Standard error	0,325	0,517
Median	19,205	20,2
Mode	/	/
Standard deviation	0,4596	1,2684
Variance	0,2112	1,6089
Kurtosis	/	–0,6372
Skewness	/	–0,0593
Range	0,65	3,37
Minimum	18,88	18,63
Maximum	19,53	22
Sum	38,41	122,89
Sample size	2	6
95% Confidence interval	4,1295	1,3311

From the table above, we can conclude that gravels are characterized with the close minimal value, but big difference in its maximal value 19,53 kN/m^3 in Zvegor-Stamer and 22,00 kN/m^3 in Lavci. Since this obvious difference, there is deviation in the average value, which is 19,21 kN/m^3 and 20,48 kN/m^3, for Zvegor-Stamer and Lavci, respectively.

Table 2. Unit weight of sands (in kN/m^3)

Sands	**Zvegor-Stamer**	**Suvodol**	**Lavci**	**Brod-Gneotino**
Average value	18,98	19,48	19,73	18,55
Standard error	0,2992	0,1788	0,1374	0,2780
Median	19,18	19,565	19,985	18,72
Mode	/	19,81	19,8	/
Standard deviation	0,7916	1,2386	0,8687	1,3331
Variance	0,6267	1,5341	0,7547	1,7772
Kurtosis	0,4486	2,1985	3,5553	4,1310
Skewness	–1,0409	0,6187	–1,3222	–1,6424
Range	2,31	6,3	4,7	6,23
Minimum	17,57	16,9	16,5	14,23
Maximum	19,88	23,2	21,2	20,46
Sum	132,83	935,14	789,27	426,76
Sample size	7	48	40	23
95% Confidence interval	0,7321	0,3596	0,2778	0,5765

Minimal value of the unit weight of the sands from the coal deposits is the biggest from 17,57 kN/m^3 for Zvegor-Stamer and the least value of 14,23 kN/m^3 for Brod-Gneotino. The maximal value varies in all coal deposits, i.e. it varies from 19,88 kN/m^3 to maximal value of 23,20 kN/m^3. The average values have small deviation, that is 18,55 kN/m^3 for Brod-Gneotino, 18,98 kN/m^3 for Zvegor-Stamer, 19,48 kN/m^3 for Suvodol and 19,73 kN/m^3 for Lavci. There is obvious difference in unit weight of sands at Suvodol and Brod-Genotino, although these coal deposits are in the same sedimentation basin.

Table 3. Unit weight of silt (in kN/m^3)

Silts	**Zvegor-Stamer**	**Suvodol**	**Lavci**	**Brod-Gneotino**
Average value	18,36	18,84	17,99	17,65
Standard error	0,29	0,0720	0,8851	0,3560
Median	18,36	18,88	17,905	18,24
Mode	/	18,94	/	17,86

Standard deviation	0,4101	0,4136	2,1682	2,4409
Variance	0,1682	0,1710	4,7009	5,9582
Kurtosis	/	3,0037	–0,8651	0,4249
Skewness	/	0,8640	0,3194	–0,8172
Range	0,58	2,13	5,91	11,97
Minimum	18,07	18,1	15,29	11,02
Maximum	18,65	20,23	21,2	22,99
Sum	36,72	621,92	107,94	829,74
Sample size	2	33	6	47
95% Confidence interval	3,6848	0,1466	2,2753	0,7167

Minimal values of the unit weight for silts are 18,07 kN/m^3 and 18,10 kN/m^3, in Zvegor-Stamer and Suvodol, respectively, 15,29 kN/m^3 in Lavci and the lowest value of 11,02 kN/m^3 in Brod-Gneotino. The maximal value is the lowest in Zvegor-Stamer with 18,65 kN/m^3 and the highest with 22,99 kN/m^3 in Brod-Gneotino. The average values are almost the same, i.e. for 17,99 kN/m^3 and 17,65 kN/m^3 for Lavci and Brod-Gneotino and higher, i.e. 18,36 kN/m^3 and 18,85 kN/m^3 for Zvegor-Stamer and Suvodol, respectively.

Table 4. Unit weight of clays (in kN/m^3)

Clays	**Zvegor-Stamer**	**Suvodol**	**Lavci**
Average value	17,81	18,37	18,73
Standard error	3,085	0,3563	0,2582
Median	17,815	18,6	19,065
Mode	/	/	18,74
Standard deviation	4,3628	0,8727	1,6330
Variance	19,0344	0,7615	2,6667
Kurtosis	/	–1,5238	2,7559
Skewness	/	–0,4795	–1,4517
Range	6,17	2,19	8,08
Minimum	14,73	17,21	13,02
Maximum	20,9	19,4	21,1
Sum	35,63	110,2	749,23
Sample size	2	6	40
95% Confidence interval	39,1986	0,9158	0,5223

It can be noticed differences in the values of minimal and maximal values of unit weight at clays in all deposits. In point of view of average values, there are

relatively close values of 18,37 kN/m^3 and 18,73 kN/m^3 in Suvodol and Lavci, and lower value of 17,82 kN/m^3 for Zvegor-Stamer.

Table 5. Unit weight of organic clays (in kN/m^3)

Organic clays	**Suvodol**	**Lavci**	**Brod-Gneotino**
Average value	18,52	16,10	16,80
Standard error	0,1522	0,5828	0,6103
Median	18,73	15,78	17,01
Mode	20,07	/	/
Standard deviation	1,1493	1,3032	1,7263
Variance	1,3209	1,6983	2,9801
Kurtosis	2,2970	–1,1197	–1,4510
Skewness	0,7957	–0,3277	–0,3404
Range	6,8	3,13	4,73
Minimum	16,04	14,3	14,18
Maximum	22,84	17,43	18,91
Sum	1055,77	80,49	134,42
Sample size	57	5	8
95% Confidence interval	0,3050	1,6181	1,4432

For organic clays, the differences in values on minimum and maximum of unit weight at all deposits are obvious. There are also differences at average values, i.e. the lowest average value of 16,10 kN/m^3 for Lavci, then average value of 16,80 kN/m^3 for Brod-Gneotino and much higher average value in Suvodol of 18,52 kN/m^3. The difference in size of the analyzed data is obvious.

Table 6. Unit weight of alevrolites (in kN/m^3)

Aleurolites	**Zvegor-Stamer**	**Lavci**
Average value	16,56	19,52
Standard error	0,2180	0,1685
Median	16,62	19,525
Mode	16,37	/
Standard deviation	1,7843	0,6305
Variance	3,1837	0,3975
Kurtosis	0,1699	–0,9202
Skewness	–0,1516	0,2934
Range	8,2	2
Minimum	13,36	18,7
Maximum	20,56	20,7

Sum	1109,84	273,26
Sample size	67	14
95% Confidence interval	0,4352	0,3640

The alevrolites (type of sediment consisting of bonded grains of silt) are present only in Zvegor-Stamer and Lavci. They have very big difference in minimal values and almost same maximal values for unit weight. In average values, also is clear difference, that is 16,56 kN/m^3 in Zvegor-Stamer and 19,52 kN/m^3 in Lavci.

Table 7. Unit weight of coals (in kN/m^3)

Coals	Zvegor-Stamer	Suvodol	Lavci	Brod-Gneotino
Average value	14,21	12,41	13,97	11,88
Standard error	0,045	0,1318	1,4989	0,3992
Median	14,215	12,54	11,14	11,99
Mode	/	/	/	/
Standard deviation	0,0636	0,6040	3,9658	0,6916
Variance	0,0040	0,3648	15,7276	0,4783
Kurtosis	/	1,2736	–1,9778	/
Skewness	/	–0,9868	0,6376	–0,6976
Range	0,09	2,47	8,6	1,37
Minimum	14,17	10,75	10,61	11,14
Maximum	14,26	13,22	19,21	12,51
Sum	28,43	260,67	97,81	35,64
Sample size	2	21	7	3
95% Confidence interval	0,5718	0,2749	3,6678	1,7180

The differences of minimal, maximal and average values of unit weight for the coal of all analyzed deposits are obvious. According the average values, lowest unit weight of 11,88 kN/m^3 has the coal in Brod-Gneotino, then 12,41 kN/m^3 has the coal in Suvodol, 13,97 kN/m^3 has the coal in Lavci and the highest unit weight of 14,22 kN/m^3 has the coal in Zvegor-Stamer.

3.3. Coefficient of filtration

The coefficient of filtration is parameter which determines the water permeability of the local environment, i.e. the deposit's sediments and can be determined by field and laboratorial experiments, as well as by empirical equations. The subject of this analysis is laboratory obtained values.

Table 8. Coefficient of filtration (*kf* [cm/s)] of sands

Sands	Zvegor-Stamer	Suvodol	Lavci	Brod-Gneotino
Average value	0,000126	0,00012	0,000842	0,002303
Standard error	$5,44\cdot10^{-5}$	$5,07\cdot10^{-5}$	0,000318	0,000492
Median	$6,26\cdot10^{-5}$	$8,16\cdot10^{-5}$	$1,65\cdot10^{-5}$	0,00078
Mode	/	$8,16\cdot10^{-5}$	0,003	0,00078
Standard deviation	0,000154	0,000209	0,001772	0,00023
Variance	$2,37\cdot10^{-8}$	$4,36\cdot10^{-8}$	$3,14\cdot10^{-6}$	0,002459
Kurtosis	–0,07193	13,47091	3,173536	$6,05\cdot10^{-6}$
Skewness	1,336927	3,526208	2,063464	–0,36507
Range	0,000383	0,000889	0,0064	0,856078
Minimum	$6,12\cdot10^{-7}$	$3,68\cdot10^{-6}$	$5,22\cdot10^{-9}$	0,0081
Maximum	0,000384	0,000893	0,0064	0,0001
Sum	0,001008	0,002043	0,026099	0,0082
Sample size	8	17	31	25
95% Confidence interval	0,000129	0,000107	0,00065	0,001015

By the results from the upper table, we can conclude that the coefficient of filtration (average value) for the sands form all coal deposits has approximately same values, which are values in the interval $(10^{-4},9\cdot10^{-4})$, except for deposit Brod-Gneotino with average values of $2\cdot10^{-3}$.We can make a conclusion that the comparative analysis is made on the base of large number of data for the coefficient of filtration for all coal deposits. Therefore, this analysis can be taken with big confidence.

Table 9. Coefficient of filtration of silts (*kf* [cm/s)]

Silts	Suvodol	Lavci
Average value	$8,61\cdot10^{-6}$	$1,98\cdot10^{-5}$
Standard error	$2,88\cdot10^{-6}$	$8,92\cdot10^{-6}$
Median	$7,21\cdot10^{-6}$	$1,03\cdot10^{-5}$
Mode	/	/
Standard deviation	$5,76\cdot10^{-6}$	$2,18\cdot10^{-5}$
Variance	$3,32\cdot10^{-11}$	$4,77\cdot10^{-10}$

Kurtosis	0,622432	−1,30632
Skewness	1,124843	0,934925
Range	$1{,}3\cdot10^{=5}$	$5{,}08\cdot10^{-5}$
Minimum	$3{,}53\cdot10^{-6}$	$1{,}7\cdot10^{-6}$
Maximum	$1{,}65\cdot10^{-5}$	$5{,}25\cdot10^{-5}$
Sum	$3{,}44\cdot10^{-5}$	0,000119
Sample size	4	6
95% Confidence interval	$9{,}17\cdot10^{-6}$	$2{,}29\cdot10^{-5}$

By the upper table, we can see that the coefficient of filtration (average value) for the silts from Suvodol and Lavci has almost same value with range in the interval $(10^{-6}, 9\cdot10^{-6})$.

Table 10. Coefficient of filtration of alevrolites (*kf* [cm/s)]

Alevrolites	**Zvegor-Stamer**	**Lavci**
Average value	$2{,}51\cdot10^{-6}$	$4{,}73\cdot10^{-8}$
Standard error	$9{,}42\cdot10^{-7}$	$1{,}15\cdot10^{-8}$
Median	$1{,}09\cdot10^{-6}$	$5{,}22\cdot10^{-8}$
Mode	/	/
Standard deviation	$2{,}98\cdot10^{-6}$	$2\cdot10^{-8}$
Variance	$8{,}88\cdot10^{-12}$	$3{,}98\cdot10^{-16}$
Kurtosis	0,293483	/
Skewness	1,206392	−1,03241
Range	$8{,}36\cdot10^{-6}$	$3{,}9\cdot10^{-8}$
Minimum	$1{,}73\cdot10^{-7}$	$2{,}54\cdot10^{-8}$
Maximum	$8{,}53\cdot10^{-6}$	$6{,}44\cdot10^{-8}$
Sum	$2{,}51\cdot10^{-5}$	$1{,}42\cdot10^{-7}$
Sample size	10	3
95% Confidence interval	$2{,}13\cdot10^{-6}$	$4{,}96\cdot10^{-8}$

By the table, easily it can be seen that the coefficient of filtration (average value) for the alevrolites from Zvegor-Stamer and Lavci has different values, the coefficient of filtration is around $n\cdot10^{-6}$ and for Lavci is $n\cdot10^{-8}$, $1\le n\le 9$.

3.4. Correlation between Unit weight (γ) –dry unit weight (γ_d)

Using all available data for the dry unit weight and unit weight from all coal deposits, correlation analysis was made. At all analysis, the coefficient of correlation is greater than $0,5$. In sequel, a correlation analysis for gravels, sands, silts, clays, organic clays, alevrolites and coals will be presented.

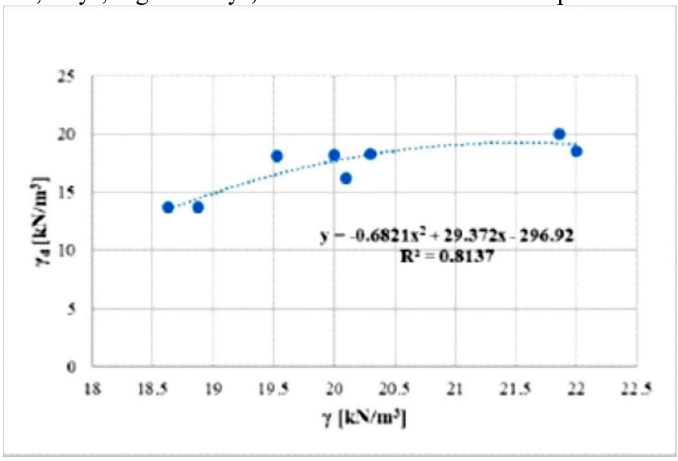

Figure 2. Correlation between unit weight and dry unit weight for gravels

From the correlation analysis for gravels, we can conclude that there is strong dependence between the parameters with coefficient of correlation ($R^2 = 0,81$) and the regression equation is given by

$$\gamma_d = -0,6821\gamma^2 + 29,372\gamma - 296,92 .$$

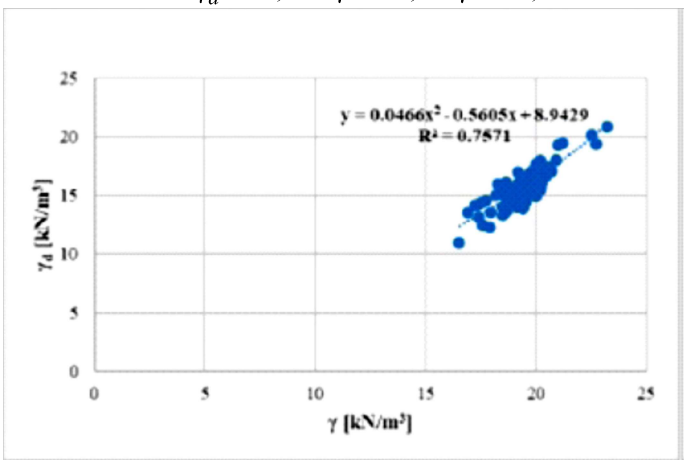

Figure 3. Correlation between unit weight and dry unit weight for sands

We can conclude that there is strong dependence for sands, between the parameters with coefficient of correlation ($R^2 = 0,76$) and regression equation given by

$$\gamma_d = 0,0466\gamma^2 - 0,5605\gamma + 8,9429 .$$

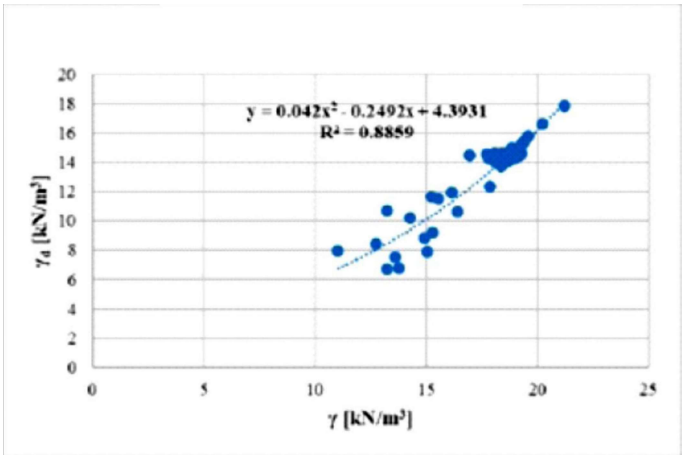

Figure 4. Correlation between unit weight and dry unit weight for silts

We can conclude that there is strong dependence for silts, between the parameters ($R^2 = 0,88$) and regression equation given by

$$\gamma_d = 0,042\gamma^2 - 0,2492\gamma + 4,3931 .$$

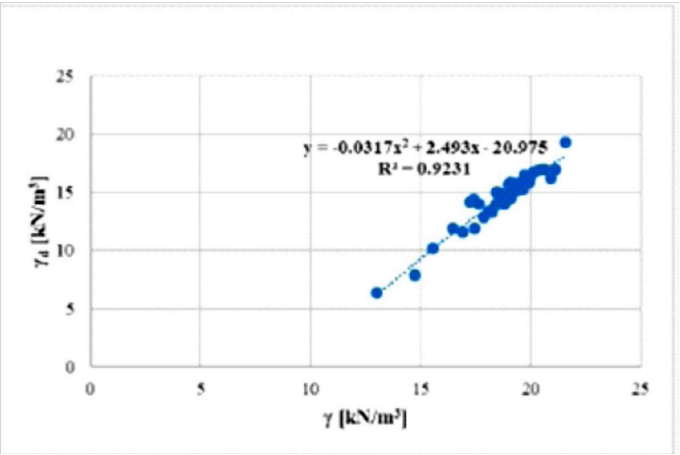

Figure 5. Correlation between unit weight and dry unit weight for clays

According to the analysis, we can conclude that for the clays there exist very strong dependence between parameters ($R^2 = 0,92$) with regression equation

$$\gamma_d = -0,0317\gamma^2 + 2,493\gamma - 20,975 .$$

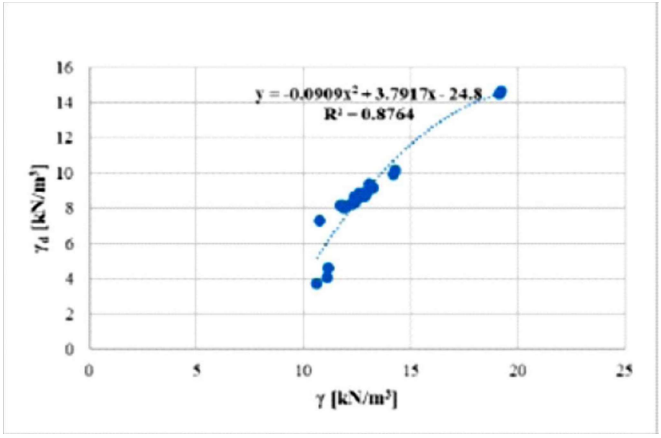

Figure 6. Correlation between unit weight and dry unit weight for coals

According to the analysis, we can conclude that there is strong dependence between parameters ($R^2 = 0,88$), and the regression is given with the equation

$$\gamma_d = -0,0909\gamma^2 + 3,7917\gamma - 24,8 .$$

3.5. Correlation between Porosity (n) – Coefficient of filtration (kf)

From the all available data for the coefficient of filtration and porosity at all coal deposits, only for the sands the regression was made. There are used twelve samples of pairs porosity and coefficient of filtration.

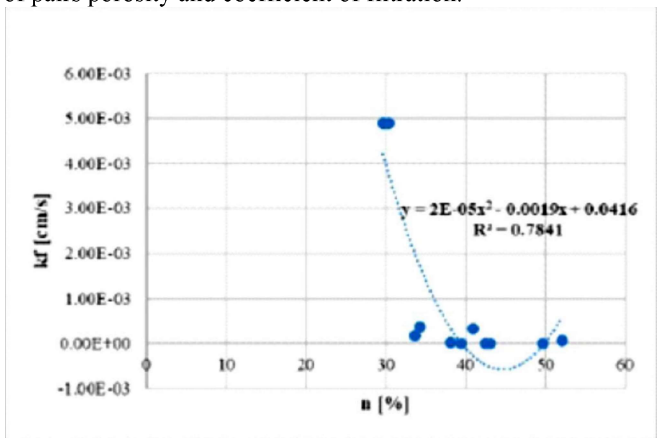

Figure 7. Correlation between porosity and coefficient of filtration for sands

We can conclude that there is strong dependence between parameters ($R^2 = 0,78$) with the regression equation

$$kf = 2 \cdot 10^{-5} n^2 - 0,0019n + 0,0416 .$$

3.6. Other attempts for correlation

Having on mind the already given dependences with coefficient of correlation greater than $0,5$ (giving the existence of some dependence between parameters), an attempt for other correlations for other geotechnical parameters was made. There was an analysis of unit weight (γ) and modulus of compressibility (Mv), but there is no correlation between these parameters, i.e. the obtained coefficient of correlation is less than $0,3$, which means that there is no apparent regression connection between these two parameters. Also, an analysis for cohesion (c) and modulus of compressibility (Mv) was made, but again the obtained result points that there is no apparent regression dependence between these two geotechnical parameters.

Another unsuccessful correlation is connections between residual angle of friction (φ_{rez}) and angle of friction (φ), as well as modulus of compressibility (Mv) and uniaxial strength (qu). In order to obtain exact and confident analysis it is very important to have large set of samples of pairs of these parameters, which in our case was absent.

4. CONCLUSIONS

Geotechnical engineering, besides in the civil engineering and environment protection, is quite important science in coal exploration, where the problems of stability of excavations, whether they are underground or surface excavations are present.

Since the surface excavation is the only way of coal exploitation in our country, it is necessary to opening of surface mines with relatively great depth, which leads to the need for design of stable slopes. In order to achieve this, is very important that detailed knowledge for geological and geotechnical profile of the considered deposit is secured.

For that purpose, the complex investigations, consisting of drilling of exploratory boreholes and field experiments, as well as laboratory tests are made. The main goal of these activities is to determine the representative geotechnical parameters for all types of materials in the deposit.

The problems given in this paper are interesting of scientific and practice point of view. The coal deposits are characterized with specific geotechnical characteristics, hence applying the approach presented in this paper give us easier determination of the geotechnical parameters, with drawing of

conclusions according to authoritative statistical parameters, correlations with big enough coefficient of correlation, rejection of no logical data, efficiency in the definition of vertical geotechnical profile of the deposits. In this way, we obtain a direction in more efficient and more economical programming of future investigations and explorations of coal deposits.

Therefore, the findings of this paper should be considered very relevant and presented techniques and results can be suggested and applied in engineering practice. The approach can be upgraded with new scientific researches and practical explorations, and thus establishing geotechnical database which will enable us to give more confident correlations applicable in practical works at coal mines.

References

[1] B. H. G. Brady, E. T. Brown, *Rock Mechanics for Underground Mining*, 3rd Edition, Chapman and Hall, London, 2006.
[2] W. J. DeCoursey, *Statistics and Probability for Engineering Applications*, College of Engineering, University of Saskatchewan, Saskatoon, 2003.
[3] *Елаборат за геомеханички истражни работи и испитувања на подинската јагленова серија во П.К. "Суводол",* Градежен институт "Македонија", а.д., Скопје, 2005.
[4] *Елаборат за изведени детални геомеханички испитувања и истражувања на јагленово наоѓалиште "Лавци",* Геохидроконсалтинг, Скопје, 2017.
[5] *Елаборат од изведени геомеханички истражувања и лабораториски испитувања во јагленово наоѓалиште "Звегор - Стамер",* Градежен институт "Македонија", а.д., Скопје, 2017.
[6] *Елаборат за изведени детални геомеханички испитувања и истражувања на јагленово наоѓалиште "Брод - Гнеотино",* Геохидроконсалтинг, Скопје, 2018.
[7] М. Георгиева, *Случајни грешки*, Математичка школа, Скопје, 1982.
[8] Ilic Lj., *Geomehanika na povrsinskim kopovima*, Savez inzinjera I tehnicara Jugoslavije, Beograd, 1991.
[9] Z. A. Ivkovic, *Matematicka statistika,* Naucna knjiga, Beograd, 1982.
[10] М. Јовановски, *Инженерска геологија,* Градежен факултет, Скопје, 2012.
[11] Р. Малчески, В. *Малческа, Теорија на веројатност,* Армаганка, Скопје, 2019.
[12] J. H. McDonald, K. W. Dunn, *Statistical tests for measures of colocalization in biological microscop,* Journal of Microscopy 252 (2013), 295-302.
[13] B. H. McCardle, *Lines, models, and errors: Regression in the field,* Limnology and Oceanography 48 (2003), 1363-1366.

[14] D. S. Moore, *The Basic Practice of statistics,* W.H. Freedman and Company, New York, 2010.
[15] I. Pavlic, *Statisticka teorija i primjena,* Tehnicka knjiga, Zagreb, 1977
[16] A. Papoulis, *Probability, Random variables and Stochastic Processes,* International Student Edition, New York, 1965.
[17] D. G. Price, M. H. De Freitas, H. R. G. K. Hack, I. E. Higginbottom, J. L. Knill, M. Maurenbrecher, In: De Freitas, M.H. (Eds.) *Engineering Geology – Principles and Practice,* Springer-Verlag, Berlin, Heidelberg, 2009.
[18] T. Lwin, J. S. Maritz, *An analysis of the linear-calibration controversy from the perspective of compound estimation,* Technometrics 24 (1982), 235-242.

Geohydroconsulting , Skopje, Macedonia
E-mail address: b_nedelkovska@hotmail.com
Faculty of Civil Engineering, Ss. Cyril and Methodius University, Skopje, Macedonia
E-mail address: pesevski@gf.ukim.edu.mk
Faculty of Civil Engineering, Ss. Cyril and Methodius University, Skopje, Macedonia
E-mail address: jovanovski@gf.ukim.edu.mk
Faculty of Civil Engineering, Ss. Cyril and Methodius University, Skopje, Macedonia
E-mail address: *velinovd@gf.ukim.edu.mk*
Faculty of Civil Engineering, Ss. Cyril and Methodius University, Skopje, Macedonia
E-mail address: *misajleski@gf.ukim.edu.mk*

DETERMINATION OF A FLOOD WAVE PROPAGATION CAUSED BY HIGH INTENSITY RAINFALLS USING PROBABILITY TECHINQUES

UDC: 519.248:556.166

Mihail Naumovski, Daniel Velinov, Zoran Misajleski

Abstract. As the most unexpected occurrence in hydrotechnical engineering, high intensity rainfalls can lead to a significant increase in water level, creating a flood wave propagation in the catchment area. Therefore, it is necessary to predict the scenario that can be critical in terms of creating a high – performed hazard, caused by the flood wave propagation, which influence can affect not only on the catchment area but also on the human beings nearby. In that case, as better the hazard is assessed, the better will the consequences be treated.

Even upon the receipitation of high intensity rainfall data base, from the hydrometeorological station, it processing can be approached. The data base is obtained as a pluviogram in every tenth minute, from which can be created a hydrogram of the maximum rainfall intensity, whose values are given in time intervals not less than the pluviograph time. So, this kind of hydrograph is the base of determination of the maximum value of the rainfall intensity that can contribute to the creation of the flood wave propagation. Also, it is important that the high intensity rainfall data base is given for a period of tenth minutes pluviograms, measured at least of a year. Hence, the created hydrogram is the base of defining the frequency of occurrence of the maximum high intensity rainfall value, which is inversely proportional by the measured period. In this way, it can be determined the variables which will be used for creating a distribution of the probability density function. The probability density function is mostly based on a Gumbel distribution, so the results are the best possible simulated. The variables are based on the parameters of a time interval sequences, which refer to the measured period.

In this case, the benefit using probability density function is not just the determination of the value of maximum intensity rainfall which causes flood wave propagation, but also is the determination of the frequency of its occurrence. This helps to correct hazard assessment in order to design buildings in hydrotechnical engineering, which are permanent and lasting, serving the surroundings, not destroying the environment.

1. INTRODUCTION

Nowadays, engineers in the domain of hydrotechnical engineering are being more aware of determinating and managing a problem, such as high intensity rainfalls, than in the past. This is due to the contemporary investigations and technologies, developed on treating such an occurrence with a stochastic character, which can cause a catastrophical scenario followed by large-scale consequences. Nevertheless there are many circumstances that can undoubtedly, as in the engineering word is said, affect to the genetic code of the buildings,

which are build just for the benefit of the mankind and the environment. And when once the genetic code is destroyed, than every idea of quality hydrotechnical engineering fades away.

The stochastic character of the high intensity rainfalls is generally due to the probability of their occurrence. Namely, there are no ways to decisively predict when and in which amount of can the high intensity rainfalls appear. This may cause distortion in the flow of watercourses and rivers, followed by increasing in water level. Such a phenomenon can inevitably form a flood wave propagation, which poses a huge threat as it is said in the abstract. For that purpose, scientific research activity till now is based on determinating the flood wave propagation, by processing the high intensity rainfalls data base, measured nearby the location in question. This location in question refers to the catchment area followed by the watercourse or the river in which the flood wave propagation is made.

2. CONDITIONS AND MEASUREMENTS

Not always in the surroundings of the location in question can a hydrometeorological station be found. Therefore, it is necessary to obtain a clear representation of the location, more specifically to the distance from the nearest hydrometeorological station. If there are more of them the relevant measurements for the high intensity rainfalls data base will be taken from all of them, but before the processing, they will be optimized and fused into one. This fusion is always made of one of well know methods in hydrotechnical engineering, especially in hydrology: *Method of Tiesen* or *Method of arithmetic mean*. Some investigations show that both methods give approximate results, but the *Method of arithmetic mean* has greater degree of confidentiality in showing the real condition of the high intensity rainfalls data base.

Figure 1. Construction of the automatical pluviograph.

In every hydrometeorological station, there are not just instruments that measure one parameter, as precipitation which is directly connected with the high intensity rainfalls, but also parameters that affect on it such as: humidity, windiness, the number of sunniest days in the year and flora and fauna nearby. The instrument which measures the precipitation is called pluviograph. There are two types of pluviographs. The traditional one, which is being used by many of the engineers from the old school and the automatical one, which is used nowadays from the new aged school engineers. In this paper it is going to be presented the automatical pluviograph, because of its widely application not just in the world, but also in our country. As it is shown in the Figure 1, the construction of the pluviograph consists of a concrete foundation, metal casing – mostly as an aluminum casing, because of its endurance of environmental aggression, in which is installed the digital device and the top construction. The digital device is connected with the funnel at the top of the construction, in which the precipitation is collecting. The digital device is tuned at a minimum time interval of ten minute, so the pluviogram could not be registered at a smaller time interval. This means that when the high intensity data base is ready for processing, it should be put a point in time, where the measurement has been stared. It is quite important to distinguish the unevenness between the started point of the measurement and the started point of the precipitation. Mostly the started point of measurement is matched with the started point of the precipitation, at the time of 00:00 when initially the day starts. As it is shown in Figure 2., the digital device takes many circumstances which can affect on the intensity of the rainfalls, so the device may register the clear image of the outside precipitation scenario.

Figure 2. Digital device of the pluviograph. (Source: https://www.pce-instruments.com)

It is also important that the pluviograph must be isolate in a radius 4-5 meters around. That is because of the digital device, which works on signals. Namely, if there is another digital device around the signals of both devices can be mixed

up and the measured data base will be incorrect and inappropriate for further processing. For proper functioning the pluviograph should be subjected to a regular inspection and maintenance.

3. DATA BASE PROCESSING AND STATISTICAL ANALYSIS

After the registered high intensity rainfalls data base is been given, the processing can be started. At first it is necessary to process the pluviograms in a table by ordering them in columns, which are represented as time interval columns. It means that the registered precipitation can be analyzed in a time frame from the smallest time interval of 10 minutes, to the largest of a day. This table has to be more organized in the detail, such as the Table 1, where are separated just the monthly maximum precipitation in a period of at least three years (the maximum period is not limited).

Table 1: Monthly maximum rainfall for characteristic time interval. (Source: [1])

Year	Month (n)	10'	20'	40'	60'	90'	120'	180'	300'	720'	1440'
2016	01	3.0	3.0	3.4	4.2	5.0	5.6	5.8	7.6	13.0	14.0
	02	1.8	2.4	3.4	3.8	4.2	4.4	5.2	5.4	5.7	7.8
	03	4.6	8.0	12.4	15.8	20.8	22.8	23.6	23.6	25.8	40.4
	04	4.0	4.8	6.4	8.0	10.0	10.2	10.8	11.8	17.0	17.6
	05	5.8	7.0	7.8	9.2	11.2	13.6	16.4	20.8	34.2	40.2
	06	7.6	8.4	10.4	12.4	12.6	13.8	18.2	22.8	22.8	22.8
	07	1.4	2.6	2.8	2.8	2.8	3.0	3.0	3.0	3.0	3.0
	08	16.0	18.6	19.2	19.2	19.2	19.2	19.4	19.8	19.8	19.8
	09	16.4	22.6	31.2	34.6	35.6	35.6	36.4	37.6	37.6	37.6
	10	2.2	2.6	3.2	3.6	4.2	4.6	6.6	8.6	9.4	9.8
	11	3.4	4.8	5.4	6.0	7.0	7.8	8.0	8.2	15.4	18.0
	12	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
2017	01	0.4	0.8	1.2	1.6	1.8	2.0	2.4	2.6	3.0	3.2
	02	0.8	1.2	1.6	2.2	2.4	2.4	2.4	3.0	3.6	3.6
	03	1.2	1.2	2.2	2.8	3.2	3.2	3.8	4.2	5.2	6.2
	04	2.6	4.8	6.0	7.2	9.0	11.6	13.4	16.8	16.8	19.0
	05	13.8	15.4	29.6	32.0	32.0	46.4	52.4	57.0	57.6	59.0
	06	3.6	4.6	4.8	6.6	8.4	8.4	10.4	10.4	10.4	20.2
	07	1.6	1.8	1.8	2.6	3.6	4.4	5.0	5.0	6.4	7.4
	08	2.0	3.0	4.2	5.0	6.6	7.6	7.2	9.4	10.0	10.0
	09	1.6	1.6	1.6	2.4	3.4	4.6	6.0	6.6	8.2	8.8
	10	1.4	2.2	4.0	4.6	5.2	6.0	6.8	8.0	10.2	10.2
	11	1.4	2.6	4.6	6.4	8.6	9.4	9.6	10.6	17.4	25.8
	12	6.2	6.2	7.6	9.4	11.0	12.8	14.0	17.6	22.4	37.6

2018	01	1.8	3.2	6.2	8.6	11.8	13.2	13.6	13.6	13.8	13.8
	02	1.8	2.2	2.4	3.4	4.2	5.2	6.6	8.8	11.6	16.0
	03	4.8	5.4	6.2	6.4	6.8	7.0	7.0	8.6	9.0	9.6
	04	2.0	3.4	4.6	5.4	6.6	6.6	6.6	6.8	7.4	7.4
	05	2.4	3.8	6.2	6.6	9.4	9.8	10.0	11.4	13.0	13.0
	06	8.6	12.8	13.2	13.2	13.4	13.6	13.8	14.2	14.2	17.4
	07	3.6	4.4	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
	08	4.6	5.6	7.4	7.8	8.0	8.0	8.0	10.2	10.6	10.6
	09	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	10	0.4	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
	11	4.2	5.6	6.4	7.0	8.6	11.0	15.6	18.6	22.4	26.4
	12	1.0	1.8	3.0	3.6	4.4	5.4	7.8	8.6	9.4	11.0

For further more detailed examination it is necessary to analyze the time interval columns as sequences, for which we should calculate some necessary statisctical parameters, in order to obtain more precise and vivid picture for the problem under investigation. In sequel, we recall some basic terms and equations from descriptive statistics. The arithmetic mean of the data sequence $x_1, x_2, ..., x_{n-1}, x_n$ is given by

$$\bar{x} = \frac{x_1 + x_2 + ... + x_{n-1} + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i ,$$

and its standard deviation is given by

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \text{ , for } n \geq 30 .$$

The standard deviation and the variance s^2 is common measure of spread about the mean as center. The standard deviation s is zero, when there is no spread and gets larger as the spread increases. The mean and standard deviation are good descriptions for symmetric distributions without outliers. Additionally, concerning the standard deviation, when $n < 30$, we use the following formula

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} ,$$

measuring more accurately the standard deviation (spread).

Next, for better analysis of the data sequences, we will use scale parameter (depends on standard deviation), given by:

$$\alpha = \frac{1,282}{s},$$

and location parameter (depends on both, arithmetic mean and standard deviation) given by:

$$\beta = \bar{x} - 0,45 \cdot s .$$

Concerning the seeking for a correlation among two variables we stress the following notations and relations. If we think that a variable x may explain or even cause changes in another variable y, we call x an explanatory variable and y a response variable.

Let we have data of an explanatory variable x and a response variable y for n.

The correlation measures the direction and strength of the linear relationship between two quantitative variables. It will be abbreviated by r. Suppose that we have data on variables x and y for n individuals. The values for the first individual are (x_1, y_1), the values for the second individual are (x_2, y_2) and so on. The arithmetic mean and standard deviation of these data sequences are $\bar{x}$ and s_x for the data sequence $x_1, x_2, ..., x_n$, $\bar{y}$ and s_y for the data sequence $y_1, y_2, ..., y_n$. The correlation r between data sequences x and y is:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \cdot \left(\frac{y_i - \bar{y}}{s_y} \right).$$

The least-squares regression line is given with the equation $y = a + bx$, with a slope:

$$b = r \frac{s_y}{s_x},$$

where s_x, s_y are standard deviations of the variable x, y, respectively and r is their correlation.

The interception a is calculated by

$$a = \bar{y} - b\bar{x},$$

where $\bar{x}$, $\bar{y}$ are means of the variables x, y, respectively.

With this regression line we obtain the predicted response y for any x.

There is a close connection between correlation and the slope of the least-squares regression line. The slope and the correlation always have the same sign.

The correlation r describes the strength of a straight-line relationship. This description takes a specific form: the square of the correlation, r^2, is the fraction of the variation in the values of y that is explained by the least-squares regression of y on x. The last one can be briefly written as:

$$r^2 = \frac{\text{variation in } y \text{ as } x \text{ pulls it along the line}}{\text{total variation in observed values of } y}.$$

From the last one, it is clear that we can always find a regression line for any relationship between two quantitative variables, but the usefulness of the line

for prediction depends on the strength of the linear relationship. Hence, r^2 is almost as important as the equation of the regression line.

Analysing the quantity of the coefficient of correlation (r^2) we can determine the strength of the regression, adopted in hydrology. We have the following grades of the strength of the linear regression: 1) $R^2 < 0,3$ there isn't any dependence; 2) if $0,3 \le R^2 < 0,5$ there is some dependence; 3) if $0,5 \le R^2 < 0,7$ there is mild dependence; 4) if $0,7 \le R^2 < 0,9$ there is strong dependence; 5) if $R^2 \ge 0,9$ there is very strong dependence.

For comprehensive approach on this statistical ideas, the interested reader may consult [5] and [7]-[11].

4. ANALYSIS USING PROBABILITY TECHNIQUES

We start this section with the following table, in which the basic statistical parameters are summarized, in each specific considered time interval.

Table 2. Statistical parameters (Source: [1])

Statistical Parameters	10'	20'	40'	60'	90'	120'	180'	300'	720'	1440'
$\bar{x}$	3.84	4.94	6.56	7.51	8.52	9.57	10.60	11.86	13.66	15.92
σ	4.03	4.98	6.94	7.50	7.71	9.21	10.03	10.81	11.42	13.12
α	0.32	0.26	0.18	0.17	0.17	0.14	0.13	0.12	0.11	0.10
β	3.36	4.48	6.08	7.05	8.11	9.13	10.17	11.45	13.28	15.54

Using regression analysis on the previously presented sequences of data, organized into certain time intervals, we can obtain the following plot given on Figure 3. The following important results (not presented of the Figure 3), concerning the regression lines, can be stated:

- For $10'$, the equation for linear regression is given by
$$y = 2,4464x + 18,677,$$
with strength of the linear regression $r^2 = 0,8709$.
- For $20'$, the equation for linear regression is given by
$$y = 2,4115x + 20,138,$$
with strength of the linear regression $r^2 = 0,865$.
- For $40'$, the equation for linear regression is given by
$$y = 2,2798x + 23,593,$$
with strength of the linear regression $r^2 = 0,8452$.
- For $60'$, the equation for linear regression is given by

$$y = 2,3318x + 22,59,$$

with strength of the linear regression $r^2 = 0,8635$.

- For $90'$, the equation for linear regression is given by

$$y = 2,483x + 19,725,$$

with strength of the linear regression $r^2 = 0,8951$.

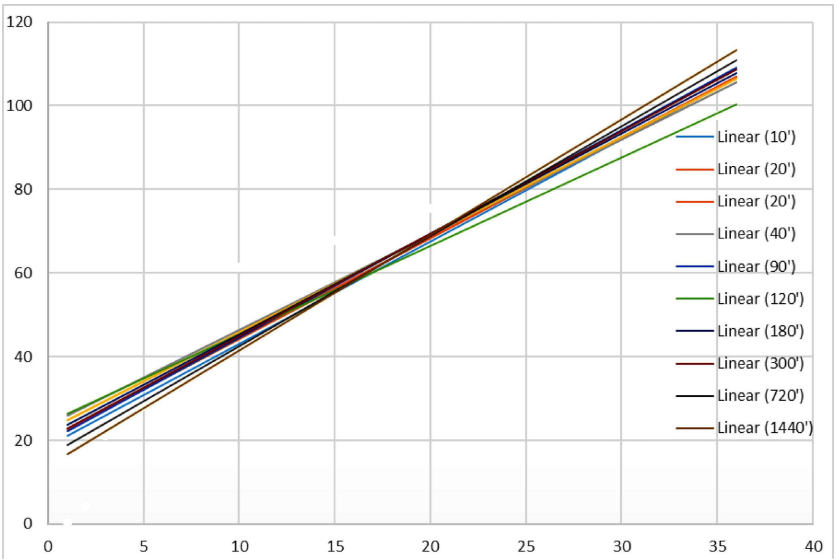

Figure 3. Regression analysis of months and maximum rainfalls for the specific time intervals

- For $120'$, the equation for linear regression is given by

$$y = 2,11x + 24,311,$$

with strength of the linear regression $r^2 = 0,6133$.

- For $180'$, the equation for linear regression is given by

$$y = 2,4024x + 31,322,$$

with strength of the linear regression $r^2 = 0,8942$.

- For $300'$, the equation for linear regression is given by

$$y = 2,4574x + 20,407,$$

with strength of the linear regression $r^2 = 0,8979$.

- For $720'$, the equation for linear regression is given by

$$y = 2,6272x + 16,363,$$

with strength of the linear regression $r^2 = 0,9306$.

- For 1440', the equation for linear regression is given by

$$y = 2,7582x + 13,926,$$

with strength of the linear regression $r^2 = 0,918$.

We can conclude that in all obtained linear regressions (except in the case of 120'), the strength of the linear regression is strong, i.e. there exists strong dependence between considered parameters. Also, we can deduce that the spread of the values is not quite big. An interesting fact that affects on a weak correlation at 120' linear regressions, is the error that occurs as a consequence of the digital deviced sleep time. Namely, every two hours per a day the digital device has to sleep not more than a minute, so the sleep time affects on a very small part of the data base not to be registered. This is most pronounced in 120' linear regressions, because it is about a punctual part of the time when the sleep time begins. Such a phenomenon does not affect on the larger time intervals, because of the significant increase of the data base.

Next, statistical parameters for certain time period (in months) and certain time intervals, or more precisely, a monthly maximum rainfall [mm] for characteristic probabilities is given.

Table 3. Statistical parameters (Source: [1])

p(z) (%)	Months (n)	10'	20'	40'	60'	90'	120'	180'	300'	720'	1440'
50	2	4.51	5.91	8.09	9.19	10.32	11.76	13.05	14.55	16.55	19.29
20	5	8.10	10.78	14.61	15.83	17.58	20.02	22.10	24.50	26.97	32.06
10	10	11.34	13.33	18.60	21.38	22.45	27.13	29.94	32.89	33.61	38.64
4	25	13.62	17.22	25.68	28.79	29.49	33.89	35.84	39.55	45.83	52.16
2	50	15.72	20.05	28.58	31.25	31.84	38.28	44.09	49.38	52.94	57.47
1	100		22.45	30.02	34.05		43.90	49.45	54.29	56.49	

On Figure 4, probability distributions of monthly maximum rainfall data i.e. the regression lines for p(z)-monthly maximum rainfalls for specific time intervals, for each considered time interval, are given.

The following facts, can be stated, concerning the obtained linear regressions. We have obtained the following:

- For 10', the equation for linear regression is given by

$$y = -0,214x + 14,399,$$

with strength of the linear regression $r^2 = 0,8911$.

- For 20', the equation for linear regression is given by

$$y = -0,2986x + 19,286,$$

with strength of the linear regression $r^2 = 0,8294$.

- For $40'$, the equation for linear regression is given by
$$y = -0,4253x + 27,097,$$
with strength of the linear regression $r^2 = 0,8511$.
- For $60'$, the equation for linear regression is given by
$$y = -0,4748x + 30,3,$$
with strength of the linear regression $r^2 = 0,8488$.

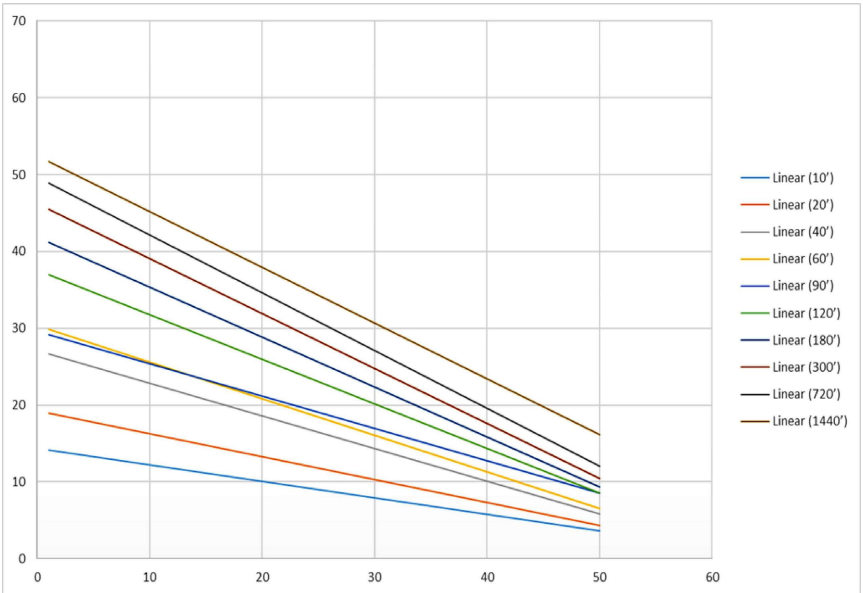

Figure 4. Probability distributions of monthly maximum rainfall data

- For $90'$, the equation for linear regression is given by
$$y = -0,42x + 29,56,$$
with strength of the linear regression $r^2 = 0,881$.

- For $120'$, the equation for linear regression is given by
$$y = -0,5806x + 37,581,$$
with strength of the linear regression $r^2 = 0,8315$.
- For $180'$, the equation for linear regression is given by
$$y = -0,6507x + 41,847,$$
with strength of the linear regression $r^2 = 0,8048$.
- For $300'$, the equation for linear regression is given by

$$y = -0,7173x + 46,262,$$

with strength of the linear regression $r^2 = 0,8026$.

- For $720'$, the equation for linear regression is given by

$$y = -0,7508x + 49,618,$$

with strength of the linear regression $r^2 = 0,8101$.

- For $1440'$, the equation for linear regression is given by

$$y = -0,7255x + 52,402,$$

with strength of the linear regression $r^2 = 0,8564$.

We can conclude that in all obtained linear regressions the strength of the linear regression is strong, i.e. there exists strong dependence between considered parameters. Such a representation of the high intensity rainfall data base, alludes to clear picture of determinating the maximum value of the precipitation that can cause a wave propagation in the catchment area. In essence, that was the purpose of this paper, not only to determinate a flood wave propagation caused by intensity rainfalls, but also to make it with a high degree of confidentiality.

5. CONCLUSIONS

Undoubtedly, the connection between engineering and mathematics can create models that can be essential in determinating processes with very small probability of occurrence. Therefore, with this paper we want to show the strength of a model that can predict a cathastrophical scenario caused by flood wave propagation, using a probability technique in order to provide protection and safety. And as an engineer challenge: The more unexpected, the better!

References

[1] B. Susinov, M. Naumovski, J. Josifovski: *Hydrological analysis of high intensity rainfalls over Topilnica tailing dam*, WMHE 2019 (16th International Symposium, Skopje, Macedonia, 2019.
[2] *PCE-FWS 20 Water Station*, www. pce-instruments.com, 05.05.2019.
[3] Z. A. Ivkovic, Matematicka statistika, Naucna knjiga, Beograd, 1982.
[4] С. Јадровски, Примена на статистички методи во хидрологијата, Магистерска работа, Градежен факултет, Скопје, 2018.
[5] Р. Малчески, В. *Малческа, Теорија на веројатност,* Армаганка, Скопје, 2019.
[6] M. Marriott, *Civil Engineering Hydraulics*, John Wiley & Sons, Inc, Singapore, 2016.

[7] J. H. McDonald, K. W. Dunn, Statistical tests for measures of colocalization in biological microscopy. Journal of Microscopy 252 (2013), 295-302.
[8] B. H. McCardle, Lines, models, and errors: Regression in the field. Limnology and Oceanography 48 (2003), 1363-1366.
[9] D. S. Moore, The Basic Practice of statistics, W.H. Freedman and Company, New York, 2010.
[10] I. Pavlic, Statisticka teorija i primjena, Tehnicka knjiga, Zagreb, 1977.
[11] A. Papoulis, Probability, Random variables and Stochastic Processes, International Student Edition, New York, 1965.
[12] Ц. Поповска, В. Ѓешовска, *Хидрологија: теорија со решени задачи*, Градежен факултет, Скопје, 2012.
[13] Ц. Поповска, С. Крстиќ, *Прирачник за реставрација на реки*, UNDP, Скопје, 2010.
[14] S. Prohaska, V. Ristić, Hidrologija kroz teoriju i praksu, Univerzitet u Beogradu, Rudarsko-geološki fakultet, Beograd, 1996.
[15] S. Prohaska, Hidrologija I, Univerzitet u Beogradu, Rudarsko-geološki fakultet, Beograd, 2006.
[16] S. Prohaska, Hidrologija II, Univerzitet u Beogradu, Rudarsko-geološki fakultet, Beograd, 2006.
[17] N. Kannan, J. P. Keating, R. L. Mason, A comparison of classical and inverse estimators in the calibration problem. Communications in Statistics: Theory and Methods 36 (2007), 83-95.
[18] T. Lwin, J. S. Maritz, An analysis of the linear-calibration controversy from the perspective of compound estimation. Technometrics 24 (1982), 235-242.
[19] Z. Shkoklevski, B. Todorovski B, *Intensity rainfalls in Republic of Macedonia*, Ss. Cyril and Methodius University, Faculty of Civil Engineering, Skopje, 1993.

Faculty of Civil Engineering, Ss Cyril and Methodius University, Skopje, Macedonia
E-mail address: naumovskimgf@outlook.com
Faculty of Civil Engineering, Ss Cyril and Methodius University, Skopje, Macedonia
E-mail address: velinovd@gf.ukim.edu.mk
Faculty of Civil Engineering, Ss Cyril and Methodius University, Skopje, Macedonia
E-mail address: misajleski@gf.ukim.edu.mk

DETERMINATION OF CERTAIN PARAMETERS IN HYDROLOGY THROUGH STATISTICS

UDC: 519.248:556.04/.06(497.7)

Violeta Gjeshovska, Daniel Velinov, Sasha Jadrovski

Abstract. Statistics uses predetermined methods for collecting and analyzing data, mainly based on probability theory. In hydrology they are used for the analysis of various hydrological and meteorological data for some past period, as well as for making decisions and conclusions about the regime in future.

The approach to statistical data processing does not have a single form in hydrology. The choice of the method depends on the accuracy or the level of confidence that this data should be analyzed i.e. interprets the course of an occurrence, or the conclusion we want to deduce for this phenomenon combined with other hydrological phenomena.

Because hydrological data are mostly limited data, statistics is the main discipline that allows obtaining complete data from the data and draws a conclusion concerning the characteristics of hydrological phenomena. The purpose of applying statistics to the hydrological variables that have been registered in the past is to determine the probability with which these variables would appear in the future.

In this paper, using old and new statistical methods, the unknown values of the annual average temperatures for five cities Skopje, Stip, D.Kapija, Prilep and Bitola in the period from 1925 to 2000 are determined. The obtained values lead us to some conclusions about their adaptability and confidentiality.

1. INTRODUCTION

Hydrology is a science for the water regime on earth surface, in atmosphere and into the soil. The processes of atmosphere's water transformation on earth surface, from the earth surface under the surface and vice versa, i.e. its mutual influence within natural environment, in every aggregate state, are main subject of investigation in this scientific field.

There are historical data for water behavior in the nature, i.e. measurements made with previously defined purpose and goal. These data form a sequence are used to obtain a certain rule for the observed phenomenon. For the measuring and observation of the parameters (biological, metrological etc.) it is formed a net of measuring stations equipped with specialized measuring equipment.

2010 *Mathematics Subject Classification.* Primary: 62P30, Secondary: 65C20, 65C60.

Key words and phrases. statistical analysis, correlation, coefficient of correlation, regression, hydrology.

The data from the monitoring net are not always available for some observed palaces or it may be available but with interceptions (deficiency) in the sequences.

Therefore, these data are not always enough; the sequences are too short or discontinuous for the analysis that we intend to do. So, in the hydrology, mathematical methods from the field of statistics and probability theory are used very often in order to fulfill (to add missing members) or continue the sequences of data. This can be done by searching a connection of at least two hydrological phenomena with measured data (sequences), such that one of them is sufficiently large and continuous.

The procedure of continuation of the sequences requests determining the strengthens of the connection between two or more hydrological phenomena (one is always with sufficiently long and continuous sequence of measured data).

The connection between hydrological and meteorological phenomena in hydrology is given with certain correlation (here under correlation we understand a stochastic connection of two or more variables).

Hydrological analysis and exact data for design of a hydrotechnical object can be made exclusively with sufficiently long and continuous sequence.

The main goal of this paper, in point of view of application, is to study the hydrological sequences of annual average temperatures at the measuring stations in Prilep, Bitola, Demir Kapiija, Stip and Skopje, to analyze the observed data, to make a decision which measuring station we should take for correlation, to fulfill the sequences, to determine and analyze the strength of the connections through coefficient of correlation and regression.

The subject of investigation in this paper is the quality of the data obtained from the measurable stations Prilep, Bitola, Demir Kapija, Stip and Skopje through statistical analysis.

The obtained data from the previously mentioned measuring stations have gaps in measuring in different time intervals. These gaps will be fulfilled with certain data obtained with linear and nonlinear correlation. For that purpose, it is made a correlation with the nearest measuring stations. These correlations are tested in order to be obtained the correlation with the biggest coefficient of correlation. In continuation, analysis of the homogeneity of the sequences of data (Kolmogorov-Smirnov test) is made and then the standard statistical parameters as arithmetic mean, standard deviation, coefficients of asymmetry and variation, coefficients of correlation and regression are determined.

There are analyzed different type of correlations (linear and nonlinear). The simple regression, i.e. dependence between two variables x and y, one dependent variable (y) and one independent variable (x) and between three variables, one dependent variable (y) and two independent variables (x_1 and x_2). The statistical testing of the obtained results is made with the standard and recommended test in these types of analysis (normalized z-test, Student's t-test, Fisher's F-test and χ^2 – test).

Analysis presented in this paper has theoretical and practical value. There given methods for determining correlation and regression in hydrology. Use of quality data for sequence continuation is very important and significant in preparation of hydrological analysis.

Application of these statistical methods can be found in many areas as ecomony [22], industry [3], thermometric [18]-[20], sociology, education, mining, agriculture, meteorology etc. The application of the statistical methods in hydrology is pretty deep and development of this field is directly dependent from usage of these methods. Since hydrology is based on observed and measured data of the investigated phenomena, processing of the obtained data is made with some statistical methods.

In regression analysis we should know at the begging which of the phenomena are independent variables, respectively dependent variables. Here the main aim is to find the form (in mathematical sense) of the connection, i.e. the formula giving the dependence of observed phenomena.

Correlation and regression are methods for describing the dependence of two or more variables. Using these methods in any field of investigation, can be stated scientifically, with certain significance, new and important results in these fields. In every case, it is important, all the influencing parameters to be taken into consideration.

2. METHODOLOGY

Hydrological sequences are series of empiric data obtained with observations and measurements of certain hydrological phenomena as rains, water levels, flows etc. The well-known parameters of a hydrological sequence are arithmetic mean $\bar{x}$, module coefficient K_i, median m, mode μ, standard quadratic deviation σ, coefficient of variation C_v, whose explanations and formulas we are going to skip, since they are very well known and basic terms in this theory. We will shortly give the more important facts about the coefficient of asymmetry, recently used in hydrology.

Let we are comparing two sequences. These two sequences can have same deviations, but different signs. Therefore, a parameter determining the degree of symmetry is introduced, known as coefficient of asymmetry. This coefficient is given by the following formula

$$C_s = \frac{\sum_{i=1}^{i=n} (K_i - 1)^3}{nC_v^3} \ .$$

For accurate calculation of this expression, it is necessary for the sequence to have at least 60 members. For shorter sequences this coefficient can be calculated empirical by the following formula

$$C_s = 2C_v \ .$$

For sequences with rare occurrence, this coefficient is given by

$$C_s = \frac{3}{6} C_v .$$

From the distribution of Pirson, we have the following

$$C_s = \frac{2C_v}{1 - K_{\min}},$$

where $(K_{\min})$ is minimal module coefficient. Using the quantity of this coefficient, it can be said the following: if $(C_s = 0)$ the sequence is symmetric; if $(0 < C_s < 0,1)$ there is no asymmetry; if $(0,1 < C_s < 0,25)$ asymmetry of the sequence is small; if $(0,25 < C_s < 0,5)$ asymmetry of the sequence is medium and if $(C_s > 0,5)$ asymmetry of the sequence is large.

When we are dealing with two or more hydrological phenomena usually exists some connection, which can be very weak to very strong. As for an example, we can state the connection between rainfall and surface leakage in a certain catchment area, although these phenomena are random in space and time. Also, there is connection on the same hydrological phenomenon measured on two near measurable stations (for example, flow of a river, measured on different profiles). The strength of this connection will be called correlation. The main task of the correlation analysis is to find a way in which one the influence of the independent variable to dependent variable is given.

The dependence between dependent variable (Y) and independent variable (X) in hydrology can be functional $Y = f(X)$, correlative $Y[Y / X]$ and stochastic.

If some certain value of the dependent variable is connected with many values of the independent variable, then this value is called a correlational value. Correlational dependences can be linear and nonlinear. Correlational dependences can be defined between dependent variable (Y) and one (X), or more independent variables (X_i). Concerning the number of variables, the correlational dependences are simple (correlation between two variables) and multiple (correlation between multiple variables). The simple correlations are formed with two random variables, one dependent Y and one independent variable X. The multiple correlation dependences are formed with one dependent variable Y and two or more independent variables X_i, where $i = 1,2,...,m$.

The correlation with one independent variable is given by

$$Y = a + b \cdot X ,$$

where $a = \overline{Y}$ and $b = \frac{S_{xy}}{S_{xx}}$.

where $\overline{X}$ is the average value of the random variable X, $\overline{Y}$ is the average value of the random variable Y, S_{xx} is the sum of the residual squares of the