Analysis Method for Determining the Suitability of Water for Human Consumption

Oleksiy Tverdokhlib [1], Denis Shavaev [1] Yurii Matseliukh [1], Aleksandr Gozhyj[2], Anna Maria Trzaskowska[3], Maksym Korobchynskyi[4], Lyubomyr Chyrun[5], and Irina Kalinina[2]

[1] *Lviv Polytechnic National University, S. Bandera Street, 12, Lviv, 79013, Ukraine*
[2] *Petro Mohyla Black Sea National University, Desantnykiv Street, 68, Mykolayiv, 54000, Ukraine*
[3] *Gdansk University of Technology, G. Narutowicza Street 11/12, 80-233, Poland*
[4] *Military Academy named after Eugene Bereznyak, 81 Y. Il'enka Str., Kyiv, 04050, Ukraine*
[5] *Ivan Franko National University of Lviv, University Street, 1, Lviv, 79000, Ukraine*

Abstract

The work establishes the main trends in determining the suitability of water for human consumption: the most common indicator of the acid-base balance of water is from 6 to 7, most of our data set are not suitable for drinking water, the most common indicator of the sulfate balance of water is from 300 to 350, the most common indicators of the carbon balance of water are within 12-15. The average and most popular value of the acid-alkaline balance of water is 7; the standard deviation from this parameter is insignificant, the indicators vary in the range of 0-14, and the sign of the acid-alkaline balance of water is quite stable.
In this work, we constructed graphs in Cartesian and polar coordinate systems, derived quantitative characteristics of descriptive statistics, and formed histograms and cumulates. Investigating this problem, we used the main methods of visualization, graphic representation and primary statistical processing of numerical data. Methods of correlation analysis of experimental data presented by time sequences were also used in work.

Keywords 1

Analysis method, determining, suitability, water, human consumption, cluster analysis, information technologies, intelligent analysis, system analysis, exponential smoothing, median filtering, data processing

1. Introduction

The problem of determining the suitability of water for human consumption belongs to the goals of sustainable development and affects the development of human capital. The study of the impact of the quality of life on the sustainable development of countries was carried out in their works [1-4]. Authors [5-7] substantiated the role of the state in the preservation of natural resources, scientists [8-10] studied the importance of existing environmental protection systems [11, 12]. Also, well-known researchers [13-15] have developed methods for assessing damages from environmental pollution and their impact on the quality of life of the population. The volumes of water bodies and their quality affect their consumption by humans [16-20]. Everyone, people use water in one form or another. Water is in food, air and, accordingly, in substances. Nowhere without water. No matter how it sounds, a person is made up of 70% water. Water ensures the body's normal functioning; therefore, any violation of the use of

MoMLeT+DS 2022: 4th International Workshop on Modern Machine Learning Technologies and Data Science, November, 25-26, 2022, Leiden-Lviv, The Netherlands-Ukraine.
EMAIL Oleksiy_Tverdokhlib@gmail.com (O.Tverdokhlib); Denis_Shavaev@gmail.com (D. Shavaev); indeed.post@gmail.com (Y. Matseliukh); alex.gozhyj@gmail.com (A. Gozhyj); atrzaskowska@gmail.com (A. M. Trzaskowska); maks_kor@ukr.net (M. Korobchynskyi); Lyubomyr.Chyrun@lnu.edu.ua (L. Chyrun); irina.kalinina1612@gmail.com (I. Kalinina)
ORCID: 0000-0002-7211-7370 (O.Tverdokhlib); 0000-0002-1707-1723 (D. Shavaev); 0000-0002-1721-7703 (Y. Matseliukh); 0000-0002-3517- 580X (A. Gozhyj); 0000-0002-0911-945X (A. M. Trzaskowska); 0000-0001-8049-4730 (M. Korobchynskyi); 0000-0002-9448-1751 (L. Chyrun); 0000-0001-8359-2045 (I. Kalinina);

water in the diet leads to inevitable consequences and even fatalities. And when there is a lot of water, but it is of dubious quality (they usually do not drink it), people start water starvation; they cannot stand and drink this water, and as a result, they get serious diseases of the digestive system, which in the absence of normal medicine (for example, Africa or poor countries) leads to deaths [16-20]. Therefore, it is very important to correctly determine whether this or that water from a certain place is suitable for consumption. Our inputs are pH, Water Hardness, Solids, Chloramines, Sulphate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity, and Potable. From the point of view of analysis, if some indicators exceed the norm too much, then even cleaning will not help here. And if the indicators are within the acceptable range, then it makes sense to attract investments, social projects, etc. [21-30]. These indicators affect the required reagents for water purification, which, in turn, affects the amount required to construct treatment facilities [31-42]. And if everything is normal, then why not inform the residents that the water is suitable for drinking, or it will be enough to boil it so as not to get an infection? And maybe this water is suitable for bottling in general; you need to remove the turbidity. Lack of water is a tragedy, especially with climate warming. When the water evaporates, it is clean, and as a result the percentage of pollution increases; add to these unscrupulous residents upstream who throw away everything they can get their hands on and we get a large-scale collapse. Therefore, this topic is more relevant than ever in the period of total pollution of the environment.

2. Related works

The most common approaches to detecting and classifying water quality were found [16-53]. You can start with works [16-20] applied a CNN-LSTM amalgam model to predict two water quality variables, dissolved oxygen and chlorophyll. The results showed that the CNN-LSTM amalgam model outperformed the separate CNN and LSTM models. Authors [16-42] compared statistical methods, including Fuzzy Logic based on modern machine learning technology and different AI methods for development of similar systems as a component of a smart city [54-65]. Inference (FLI) and WQI for water quality assessment in the community of Ikare, Nigeria [35]. They identified moderate and poor water quality using FLI and WQI methods, respectively. They also found that the FLI method was superior to the WQI method because of the relationship between the measured and standard WQI values [43-53]. To estimate dissolved oxygen in aquaculture, authors [43] proposed a synthetic model. Although CNN-LSTM models and Sparse - autoencoder - LSTMs showed excellent performance because they only predicted DO and chlorophyll, it can be difficult to deal with more water quality variables using such models. In another study [44] authors applied Extra Tree Regression (ETR), which combines multi-week studies to predict WQI values in Tsuen River, Hong Kong. They applied the ETR method to ten water quality variables. The results showed that the ETR method achieved 98% prediction accuracy, outperforming other state-of-the-art models such as support vector regression and decision trees. A complete study on the application of methods for river water quality modelling was conducted by authors [45], where they reviewed 51 articles published between 2000 and 2016. According to this study, artificial neural networks and wavelet neural networks were the most widely used methods for water quality prediction. In addition, scientists [46] developed an artificial neural network. For this study, the most significant water quality parameters were found using spatial discriminant analysis (SDA). But in another study [16] these studies can barely show an accuracy of 71%. In the work [37] applied an artificial neural network to predict WQI in the Akaki River in Ethiopia. In this analysis, an artificial neural network with eight hidden layers and 15 hidden neurons predicts WQI with more than 90% accuracy. Also, authors [47] applied an artificial neural network with one hidden layer to predict the sustainability of water quality in São Paulo, Brazil. Applying neural networks for WQI prediction requires a large amount of water quality data, which is expensive and time-consuming. Researchers [41] applied a decision tree to classify water quality status in the Klang River, Malaysia. They considered three scenarios where; they used six water quality variables in the first scenario. They then removed water quality parameters such as NH3-N, pH, and SS during each procedure to evaluate the ability of the decision tree algorithm in different situations. They achieved classification accuracies of 84.09%, 81.82%, and 77.27% in each scenario, which are higher than the 75% classification accuracy comparison [39-41]. This study used 22 water quality samples, making the model computationally expensive.

3. Methods

To solve the problems to be considered in this paper, we will use several standard methods, such as:

1. The moving average method [66, 67]. This method estimates the average level for a certain period. The longer the time interval to which the average belongs, the more smoothly the level will be smoothed, but the less accurately the trend of the original dynamics series will be described [68-70]. The moving average method is the simplest way of smoothing empirical curves. The essence of this method consists of replacing the indicator's actual values with their averaged values, which have a much smaller variation than the original levels of the series.

Moving averages calculated for odd and even numbers of time intervals are distinguished depending on the averaging period [71, 72]. A more complex calculation scheme is used in cases where an even number of elements determines the moving average. The following algorithm is used for calculation.

First, it is necessary to determine the length of the smoothing interval l, which includes l consecutive levels of the series ($l < n$) [73-75]. At the same time, it should be taken into account that the wider the smoothing interval, the greater the mutual fluctuations, and the trend of development has a smoother, smoother character. The stronger the oscillation, the wider the smoothing interval should be. Next, it is necessary to break down the entire period of observation at the site while the smoothing interval, as it were, slides along the row with a step equal to l. Calculate the arithmetic mean of the levels of the series forming each section. Replace the actual values of the row in the centre of each plot with the corresponding average values. The algorithm for calculating a simple moving average is as follows [76-79]. The definition of the moving average in the case of an even number of levels in the moving interval is complicated by the fact that then the average should be attributed only to the middle between two moments located in the middle of the smoothing interval and at such a moment no observations were made. If the graphic representation of the time series resembles a straight line, then the moving average does not distort the dynamics of the studied phenomenon.

2. Weighted moving average method [66-70, 73, 80] A more subtle technique, based on the same idea as simple moving averages, is to use weighted moving averages. If, when applying a simple moving average, all levels of the series are recognized as equal, then when calculating the weighted average, each level within the smoothing interval is assigned a weight that depends on the distance measured from the given level to the middle of the smoothing interval. When building a weighted moving average on each active site, the value of the main level is replaced by the calculated one, calculated according to the formula of the weighted arithmetic average. In other words, a weighted moving average differs from a simple moving average because the levels included in the averaging interval are summed with different weights. A simple moving average takes into account all the series levels included in the smoothing interval with equal weights, and the weighted average assigns to each level a weight that depends on the distance of the given level to the level standing in the middle of this interval [66-70, 73, 81-82]. This is because for a simple moving average in the smoothing interval, calculations are performed based on a straight line - a polynomial of the first order, and for smoothing with a weighted moving average, polynomials of higher orders, preferably of the second or third order, are used. Therefore, the simple moving average method is possibly considered a special case of the weighted moving average method [66-72]. The calculation of the moving average is presented as a simple and safe operation with a completely clear meaning. However, this operation transforms the dynamic series to a greater extent than it seems at first glance. So, if the levels of the series were independent before the smoothing, then after this transformation, the successive calculated levels (within the smoothing interval) are somewhat dependent on each other. Indeed, each level of the smoothed series has a common part with several previous and subsequent members. The algorithm of smoothing with a weighted moving average with the size of the "window" - the smoothing interval $w = 2k + 1$, which is successively shifted along the series levels and averages the levels covered by it. The formula for calculation [66-72, 83-85]:
3. Correlation field [67, 86-88]. A correlation field is a graph that establishes a relationship between variables, where X of each corresponds to the abscissa value and Y to the ordinate value of a specific unit of observation. The number of points on the graph corresponds to the number of observation units. The placement of points shows the presence and direction of

communication. To build a correlation field, you usually need to take the following steps: choose two variables that change over time. Then the value of the dependent variable is measured. As a result, the result is entered in the table. Then a coordinate grid is built, the value of the independent variable is indicated on the X axis, and the dependent variable is indicated on the Y axis. After that, you need to mark the points of the correlation field. On the X-axis for the first value of the independent variable, mark the point on the Y-axis corresponding to the value of the dependent variable. The obtained result is called the correlation field. Next, it is necessary to analyze the schedule and form a conclusion[67, 86-89].

   a. Correlation coefficient.
   b. Correlation relationship.
   c. Correlation matrix.
   d. Autocorrelation.

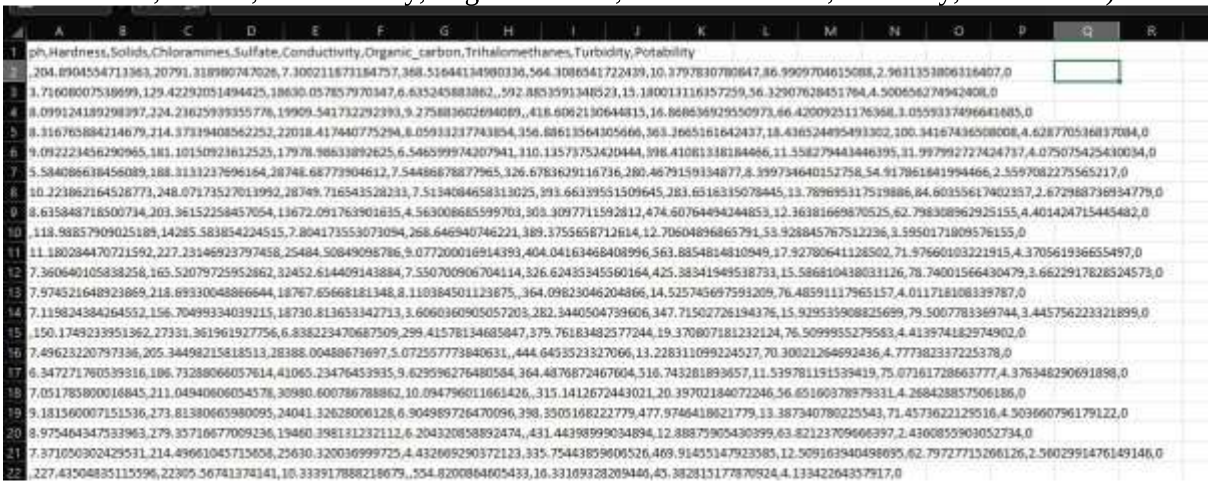
4. Cluster analysis is one of the methods of multivariate statistical analysis; that is, each observation is represented not by a single indicator but by a set of values of various indicators [5, 86, 91-99]. It includes algorithms with the help of which the clusters' formation and the distribution of objects by clusters are carried out. Cluster analysis, first of all, solves the problem of adding structure to the data and also ensures the selection of groups of objects, that is, looks for the division of the population into areas of accumulation of objects. Cluster analysis allows you to consider fairly significant volumes of data, sharply shorten and compress them, make them compact

4. Experiments

4.1. Analysis of existing software products

To begin with, we downloaded the dataset [89] and began familiarization with it.

Fig.1 is what the original dataset looks like in Excel (our dataset are pH, Water hardness, Solids, Chloramines, Sulfate, Conductivity, Organic carbon, Trihalomethanes, Turbidity, and Potable):

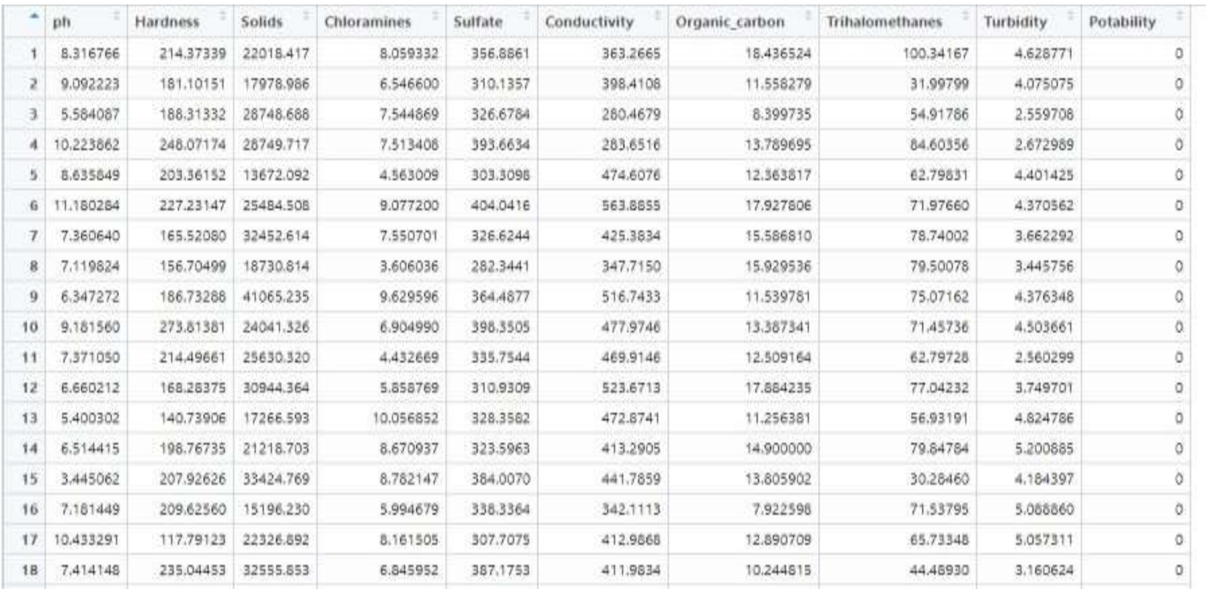


```
1  ph,Hardness,Solids,Chloramines,Sulfate,Conductivity,Organic_carbon,Trihalomethanes,Turbidity,Potability
2  ,204.8904554713363,20791.318980747026,7.300211873184757,368.5164413498036,564.3086541722439,10.3797830780847,86.9909704615088,2.9631353806316407,0
3  3.71608007538699,129.42292051494425,18630.057857970347,6.635245883862,,592.8853591348523,15.180013116357259,56.32907628451764,4.500656274942408,0
4  8.099124189298397,224.23625939355776,19909.541732292393,9.275883602694089,,418.6062130644815,16.868636929550973,66.42009251176368,3.0559337496641685,0
5  8.316765884214679,214.37339408562252,22018.417440775294,8.059332377743854,356.88613564305666,363.2665161642437,18.436524495493302,100.34167436508008,4.628770536837084,0
6  9.092223456290965,181.10150923612525,17978.98633892625,6.546599974207941,310.13573752420444,398.41081338184466,11.558279443446395,31.997992727424737,4.075075425430034,0
7  5.584086638456089,188.3133237696164,28748.68773904612,7.544868788877965,326.6783629116736,280.4679159334877,8.399734640152758,54.917861841994466,2.5597082275565217,0
8  10.223862164528773,248.07173527013992,28749.716543528233,7.5134084658313025,393.66339551509645,283.6516335078445,13.789695317519886,84.60355617402357,2.672988736934779,0
9  8.635848718500734,203.36152258457054,13672.091763901635,4.563008685599703,303.3097711592812,474.60764494244853,12.36381669870525,62.798308962925155,4.401424715445482,0
10 ,118.98857909025189,14285.583854224515,7.804173553073094,268.646940746221,389.3755658712614,12.70604896865791,53.928845767512236,3.5950171809576155,0
11 11.180284470721592,227.23146923797458,25484.50849098786,9.077200016914393,404.04163468408996,563.8854814810949,17.92780641128502,71.97660103221915,4.370561936655497,0
12 7.360640105838258,165.52079725952862,32452.614409143884,7.550700906704114,326.62435345560164,425.38341949538733,15.586810438033126,78.74001566430479,3.6622917828524573,0
13 7.974521648923869,218.69330048866644,18767.65668181348,8.110384501123875,,364.09823046204866,14.525745697593209,76.48591117965157,4.011718108339787,0
14 7.119824384264552,156.70499334039215,18730.813653342713,3.6060360905057203,282.3440504739606,347.71502726194376,15.929535908825699,79.50077833697744,3.445756223321899,0
15 ,150.1749233951362,27331.361961927756,6.838223470687509,299.41578134685847,379.76183482577244,19.370807181232124,76.50999552795583,4.413974182974902,0
16 7.49623220797336,205.34498215818513,28388.00488673697,5.072557773840631,,444.6453523327066,13.228311099224527,70.30021264692436,4.777382337225378,0
17 6.347271760539316,186.73288066057614,41065.23476453935,9.629596276480584,364.4876872467604,516.743281893657,11.539781191539419,75.07161728663777,4.376348290691898,0
18 7.05178580016845,211.04940606054578,30980.600786788862,10.094796011661426,,315.1412672443021,20.39702184072246,56.65160378979331,4.268428857506186,0
19 9.181560007151536,273.81380665980095,24041.32628006128,6.904989726470096,398.3505168222779,477.9746418621779,13.387340780225543,71.4573622129516,4.503660796179122,0
20 8.975464347533963,279.35716677009236,19460.398131232112,6.204320858892474,,431.44398999034894,12.88875905430399,63.82123709666397,2.4360085590305274,0
21 7.371050302429531,214.49661045715658,25630.320036999725,4.432669290372123,335.75443859606526,469.91455147923585,12.509163940498695,62.79727715266126,2.5602991476149146,0
22 ,227.43504835115596,22305.56741374141,10.333917888218679,,554.8200864605433,16.33169328269446,45.382815177870924,4.13342264357917,0
```

Figure 1. A selected dataset in excel

pH is an important parameter for assessing the acid-alkaline balance of water. Water hardness is mainly due to calcium and magnesium salts [90]. These salts dissolve from geological deposits through which water moves. Solids - a wide range of inorganic and organic minerals or salts, such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates, etc., can dissolve in water. This is an important parameter for water use. Chloramines are the main disinfectants used in public water systems. Sulfates are naturally occurring substances found in minerals, soil and rocks. They are present in atmospheric air, underground water, plants and food products. Conductivity: Pure water is not a good conductor of electricity but a good insulator [90]. An increase in ion concentration increases the electrical conductivity of water. Total organic carbon in source waters comes from decaying natural organic matter and synthetic sources. Trihalomethanes are chemicals found in chlorinated water. The turbidity of water depends on the number of suspended solids. Potable indicates whether the water is safe for human consumption, where one means potable and 0 means non-potable [90]. Next, we loaded our dataset into the RStudio development environment:

```
water <- read.csv( file ='D:/water_potability.csv')
```



	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
1	8.316766	214.37339	22018.417	8.059332	356.8861	363.2665	18.436524	100.34167	4.628771	0
2	9.092223	181.10151	17978.986	6.546600	310.1357	398.4108	11.558279	31.99799	4.075075	0
3	5.584087	188.31332	28748.688	7.544869	326.6784	280.4679	8.399735	54.91786	2.559708	0
4	10.223862	248.07174	28749.717	7.513408	393.6634	283.6516	13.789695	84.60356	2.672989	0
5	8.635849	203.36152	13672.092	4.563009	303.3098	474.6076	12.363817	62.79831	4.401425	0
6	11.180284	227.23147	25484.508	9.077200	404.0416	563.8855	17.927806	71.97660	4.370562	0
7	7.360640	165.52080	32452.614	7.550701	326.6244	425.3834	15.586810	78.74002	3.662292	0
8	7.119824	156.70499	18730.814	3.606036	282.3441	347.7150	15.929536	79.50078	3.445756	0
9	6.347272	186.73288	41065.235	9.629596	364.4877	516.7433	11.539781	75.07162	4.376348	0
10	9.181560	273.81381	24041.326	6.904990	398.3505	477.9746	13.387341	71.45736	4.503661	0
11	7.371050	214.49661	25630.320	4.432669	335.7544	469.9146	12.509164	62.79728	2.560299	0
12	6.660212	168.28375	30944.364	5.858769	310.9309	523.6713	17.884235	77.04232	3.749701	0
13	5.400302	140.73906	17266.593	10.056852	328.3582	472.8741	11.256381	56.93191	4.824786	0
14	6.514415	198.76735	21218.703	8.670937	323.5963	413.2905	14.900000	79.84784	5.200885	0
15	3.445062	207.92626	33424.769	8.782147	384.0070	441.7859	13.805902	30.28460	4.184397	0
16	7.181449	209.62560	15196.230	5.994679	338.3364	342.1113	7.922598	71.53795	5.088860	0
17	10.433291	117.79123	22326.892	8.161505	307.7075	412.9868	12.890709	65.73348	5.057311	0
18	7.414148	235.04453	32555.853	6.845952	387.1753	411.9834	10.244815	44.48930	3.160624	0

Fig. 2. Dataset view in RStudio

Present a graphical presentation of the dataset.

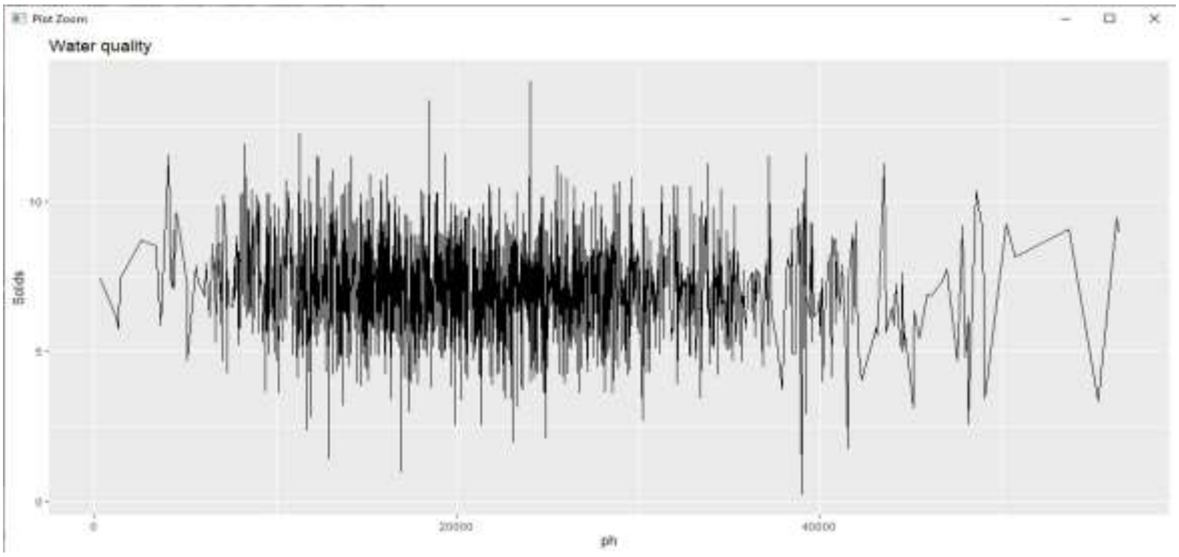


Fig. 3. Graph of dependence of ph level on Solids in water in Cartesian coordinates

For visualization, we will use the ggplot2 library, which allows you to build beautiful graphs. First, install the library:

```
install.packages ("ggplot2")
```

Program code for plotting a graph of the dependence of the degree of acidity on solids in the usual Cartesian coordinates:

```
library (ggplot2) plot1 <- ggplot () + geom_line ( aes (y = ph , x = Solids ), data = water ) plot1 + labs ( title = " Water quality ", x = " ph ", y = " Solids ")
```

The polar coordinate system is most often used for pie charts, which are bar charts stacked in polar coordinates. To write the software code for plotting the graph of the degree of acidity versus organic carbon, we used coord_polar ():

```
plot2 <- ggplot ( water , aes (x = ph , fill = Organic_carbon )) + geom_histogram ( binwidth = 15, boundary = -7.5) +
coord_polar () + scale_x_continuous ( limits = c(0,360)) plot2 + labs ( title = " Water quality ", x = " ph ", y = "
Organic_carbon ")
```

Figure 4 shows the dependence of ph on organic_carbon in polar coordinates.

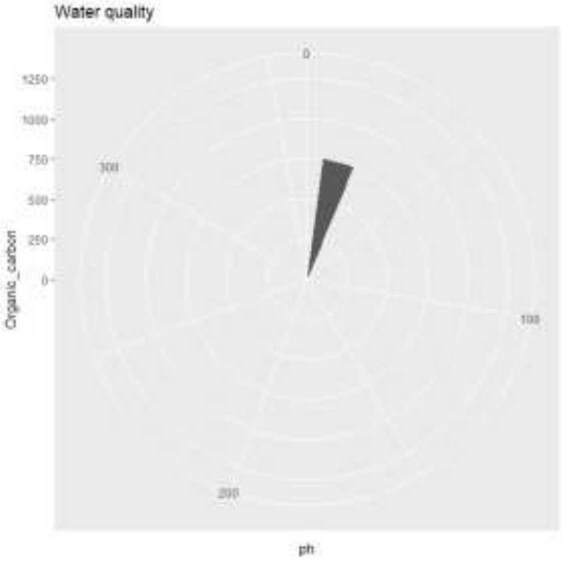


Fig. 4. Graph of dependence of ph level on Organic_carbon in water in polar regions coordinates

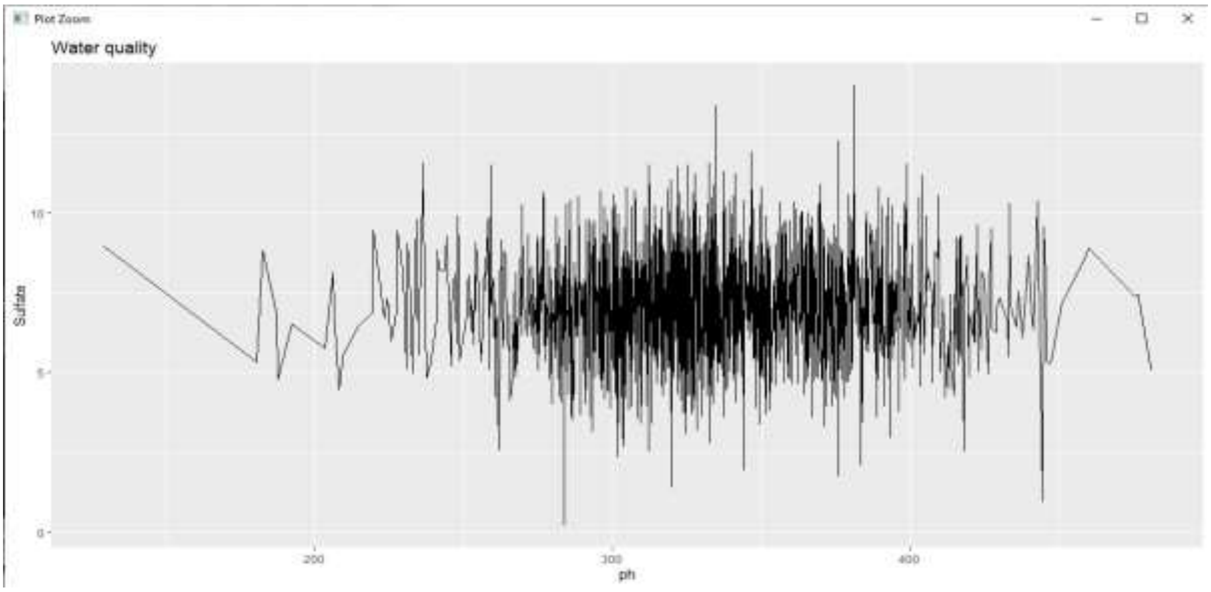


Fig. 5. Graph of dependence of ph and sulfate in Cartesian coordinates

Water acidity to sulfate content:

```
water_sorted_ph <- water [ order ( water$ph ), ]
plot1 <- ggplot () + geom_line ( aes (y = ph , x = Sulfate ), data = water_sorted_ph ) plot1 + labs ( title = " Water quality ", x =
" ph ", y = " Sulfate ")
```

A histogram is a way of graphically presenting tabular data and their distribution. A histogram can be created using the hist () function in the R programming language. This function accepts a vector of values for which the histogram is constructed.

This graph shows the dependence of ph (acidity) on solids. You can see that most of the data ranges from 15000 to 30000 for ph and 5 to 10 for solids. It can be concluded that most of the water from this dataset is not of the best quality, and in some places, it is very toxic.

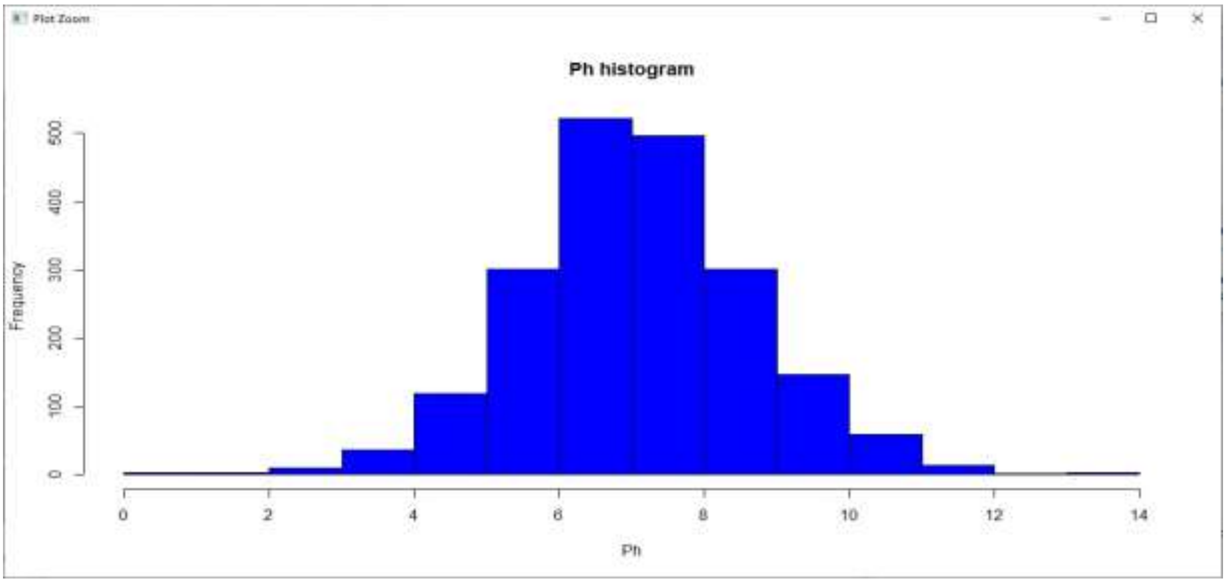


Fig. 6. ph indicator

Program code for constructing a histogram of water acidity:

```
library (ggplot2)
hist ( water$ph , main =" Ph histogram ", xlab =" Ph ", col =" blue ")
```

Similarly, the program code for building a histogram of water hardness:

```
hist ( water$Hardness ,          main =" Hardness    histogram ",           xlab =" Hardness ", col =" blue ")
```

It can be concluded that most of the water is not suitable for consumption because the indicators are too high. This histogram shows that the largest number of cases is in the interval 6-8, with about 500 cases in the interval 6-7. From Fig. 7, it can be seen that 1200 (60% of the entire sample) cases are unsuitable for use, and 800 are suitable.

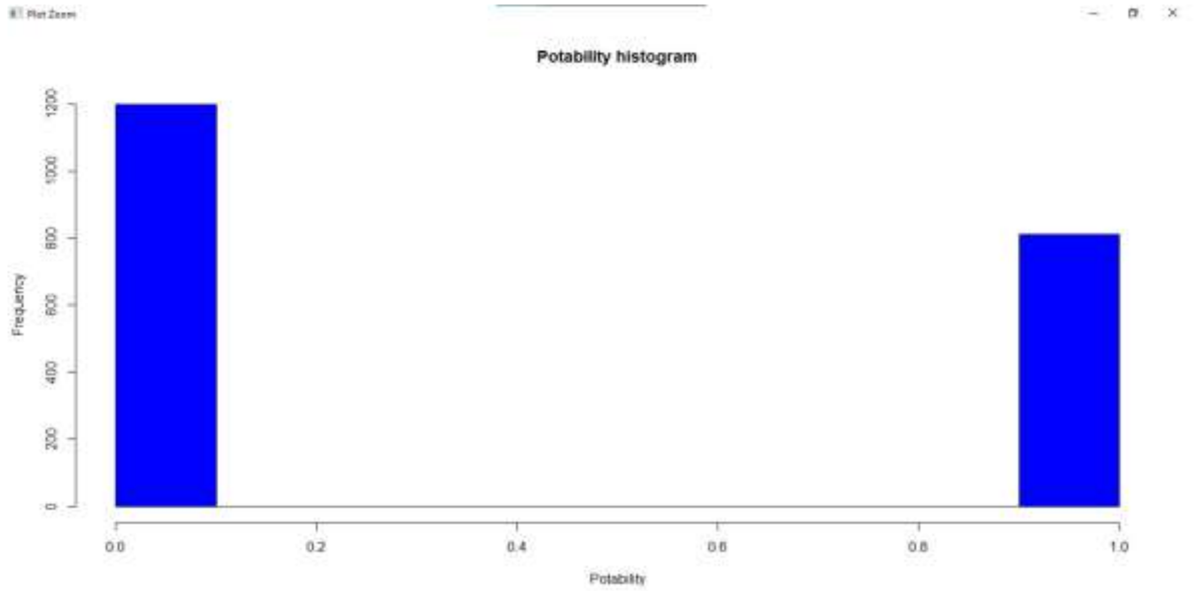


Fig. 7. Potability indicator histogram

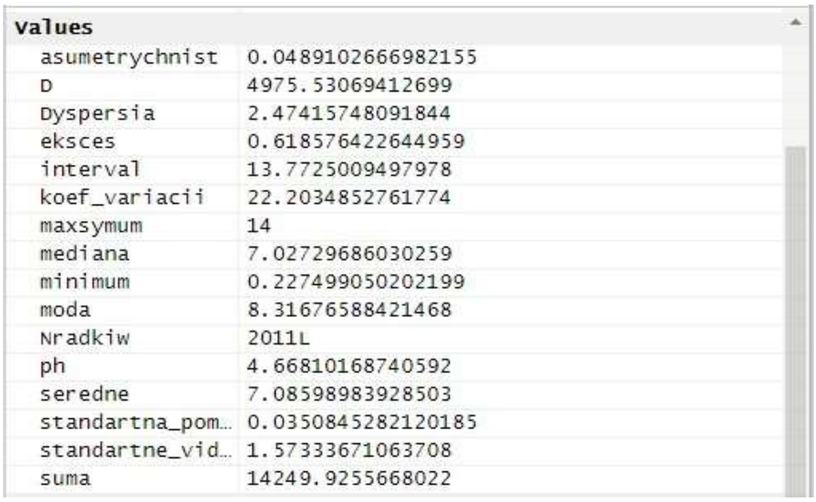
PerformanceAnalytics is a package of econometric functions for analyzing the performance and risks of financial instruments or portfolios. Let's try to determine some parameters of the pH indicator:

Arithmetic means - the average value of the sample. Let's use the mean () method :

```
library ( PerformanceAnalytics ) #Arithmetic mean seredne <- mean ( water$ph )
```

The median is the number that divides the set of sample numbers in half. Him median () method :

```
#median
median <- median ( water$ph )
```



```
Values
  asumetrychnist    0.048910266982155
  D                 4975.53069412699
  Dyspersia         2.47415748091844
  eksces            0.618576422644959
  interval          13.7725009497978
  koef_variacii     22.2034852761774
  maxsymum          14
  mediana           7.02729686030259
  minimum           0.227499050202199
  moda              8.31676588421468
  Nradkiw           2011L
  ph                4.66810168740592
  seredne           7.08598983928503
  standartna_pom…   0.0350845282120185
  standartne_vid…   1.57333671063708
  suma              14249.9255668022
```

Fig. 8. Research results

Obtained results:

- Average - 7.08599
- The standard error is 0.035
- Median - 7.027297
- Fashion - 8.316766
- The standard deviation is 1.573337
- Sample variance - 2.474157 ● Skewness - 0.6185764
- Asymmetry - 0.04891027
- Interval - 13.7725
- The minimum is 0.23
- The maximum is 14
- The amount is 14249.93
- Volume (quantity) - 2011
- Coefficient of variation - 22.2%

A more detailed analysis of the data can be found in the Discussion section.

Standard error is the deviation of the sample from the actual mean. Let's use the std() method :

```
#standard error
std <- function (x) sd (x)/ sqrt ( length (x)) standartna_pomylka <- std ( water$ph )
```

Mode is the number that occurs most often in the sample. Let's use the function getmode():

```
#fashion
getmode <- function (v) {
uniqv <- unique (v)
uniqv [ which.max ( tabulate ( match (v, uniqv )))]
}
mode <- getmode ( water$ph )
```

Standard deviation is the amount of spread relative to the arithmetic mean. To search, we use the sd() method :

```
deviation standartne_vidchylennya <- sd ( water$ph )
```

Variance is an estimate of the theoretical variance of the distribution based on the sample:

```
#dispersion D <- 0
for ( ph in water$ph ) {
D <- D + ( ph-mean ( water$ph ))**2
}
Dyspersia <-(D / nrow ( water ))
```

Skewness is a parameter that reflects the height of the distribution. We will use the moments library and the kurtosis () method :

```
#kurtosis
excess <- kurtosis ( water$ph )
```

Asymmetry reflects the skewness of the distribution relative to the mode.

skewness () method :

```
#asymmetry
asumetrychnist <- skewness ( water$ph )
```

The interval is the difference between the minimum and maximum value of the sample:

```
#interval
interval <-( max ( water$ph ) - min ( water$ph ))
```

Minimum - the smallest value of the sample

```
#minimum
minimum <- min ( water$ph )
```

Maximum - the largest sample value:

```
#maxymum maxsymum <- max ( water$ph )
```

Sum of all sample values:

```
#sum suma < - sum ( water$ph )
```

Total number of columns with data:

```
#sample size
Nradkiw <- nrow ( water )
```

The coefficient of variation is an indicator that determines the percentage ratio of the average deviation to the average value:

```
#coefficient of variation
coef_variacii <-( sd (( water$ph )) / mean (( water$ph )) * 100)
```

Cumulants are a representation of the distribution in the form of a curve, the ordinates of which are proportional to the accumulated frequencies of the variation series. To make a series of accumulated frequencies, you need to add the frequency of the second class to the frequency of the first, smallest class, then add the frequency of the third class, etc.

Cumulative sometimes have an advantage over the variation curve.

```
ph = water$ph breaks = seq (0, 14, by =0.1) ph.cut = cut ( ph , breaks , right =FALSE) ph.freq = table ( ph.cut ) cumfreq0 =
c(0, cumsum ( ph.freq ))
plot ( breaks , cumfreq0, main =" ph ", xlab =" ph ", ylab =" Number")
lines ( breaks , cumfreq0)
Hardness = water$Hardness breaks = seq (74, 315, by =1)
Hardness.cut = cut ( Hardness , breaks , right =FALSE) Hardness.freq = table ( Hardness.cut ) cumfreq0 = c(0, cumsum
(Hardness.freq ))
plot ( breaks , cumfreq0, main =" Hardnes ", xlab =" Hardness ", ylab =" Number") lines ( breaks , cumfreq0)
```

A cumulant is a continuous curve graphically depicted in a coordinate system, where the value of the characters or the limits of its intervals is indicated on the abscissa axis, and the increasing sum of frequencies is indicated on the ordinate axis.

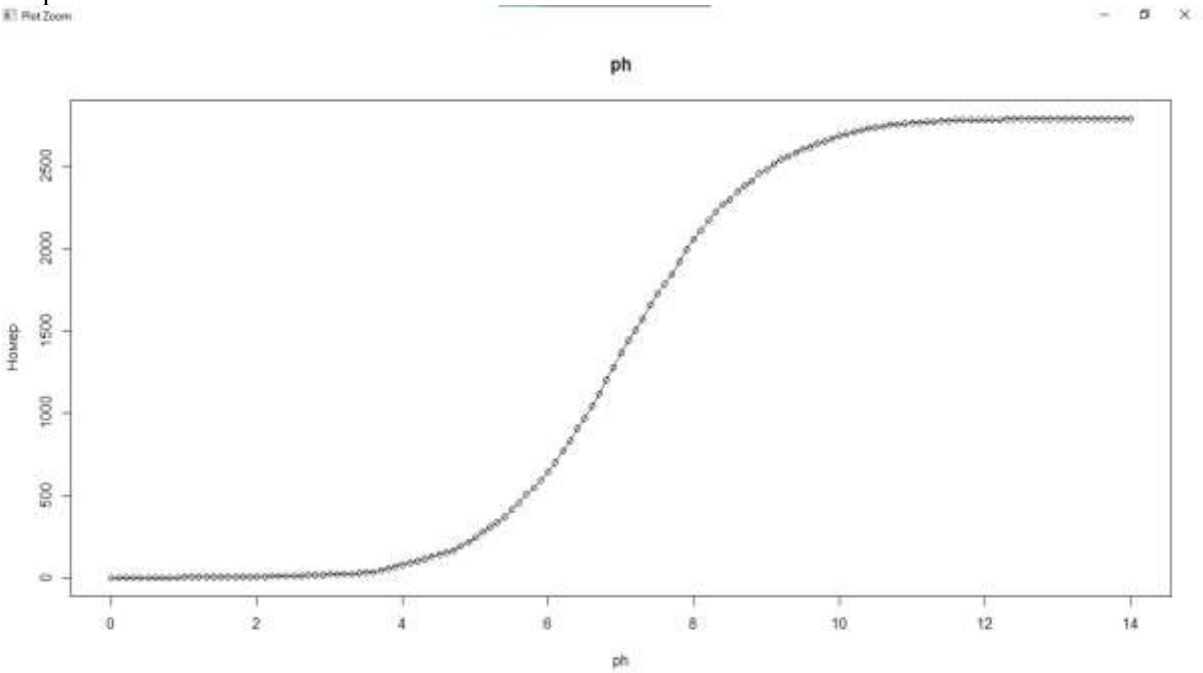


Fig. 9. Cumulative indicators of ph

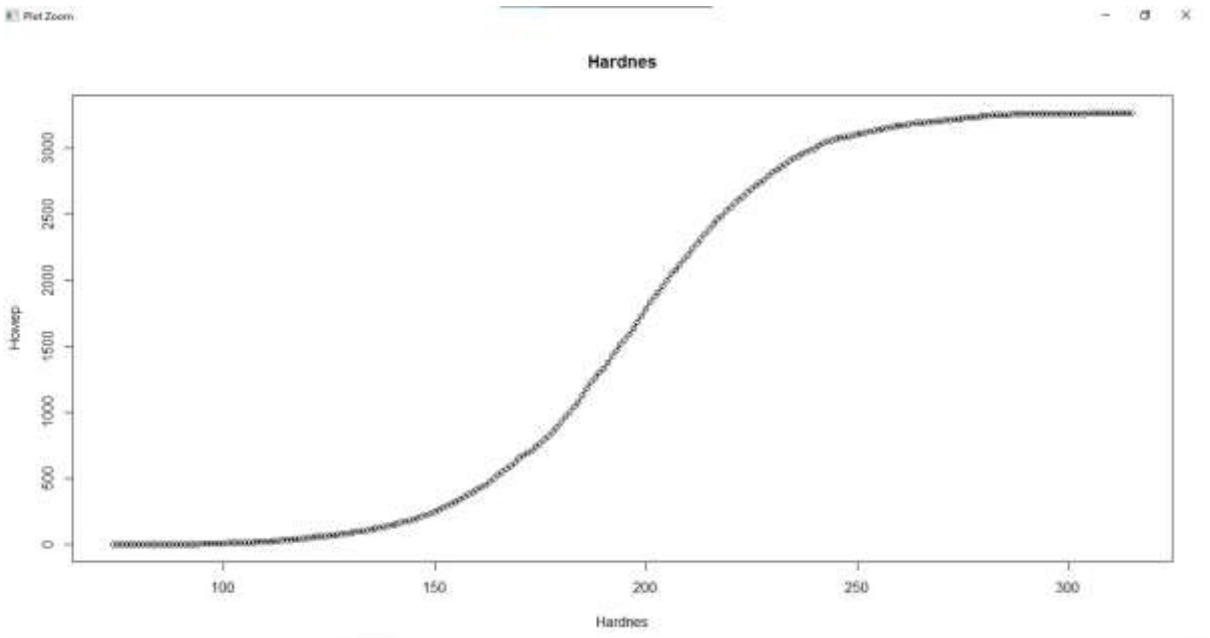


Fig. 10. Hardness indicator

Having analyzed the created cumulates, we can conclude that all indicators have a sharp increase in pollution, which correlates with the increase in the numerical value of the parameters, that is, more water with higher indicators.

5. Results and discussion

Smoothing methods reduce the influence of the random component (random fluctuations) in time series [66-90]. They make it possible to obtain more "pure" values, which consist only of deterministic components. Some methods aim to highlight only some components, for example, a trend. We will perform smoothing using different methods. We will use the following libraries:

- library (tidyverse);
- library (lubridate);
- library (fpp2);
- library (zoo);
- library (pastecs);
- library (TTR).

We import and number the data:

```
water <- read.csv( file ='D:/water_potability1.csv') id <- c(1:3276) water <- cbind ( id , water )
```

1. The moving average method [67]. We will use Kendel's formulas for smoothing according to the moving average. The method is often used for statistical evaluation in statistical hypothesis testing to determine whether two variables can be considered statistically dependent. Under the null hypothesis of independence of X and Y, the sampling distribution τ has an expected value of zero. The exact distribution cannot be characterized in terms of joint distributions but can be calculated for small samples; for larger samples, it is common to use the approximation for a normal distribution with a mathematical expectation equal to zero and a random variable variance. We will smooth our data by the following sizes of the smoothing interval w = 3, 5, 7, 9, 11, 13, 15 to obtain seven bars using the rollmean () function:

```
ma <- water %>% select ( id , Hardness ) %>% mutate (ma1 = rollmean ( Hardness , k = 3, fill = NA), ma2 = rollmean ( Hardness , k = 5, fill = NA), ma3 = rollmean ( Hardness , k = 7, fill = NA), ma4 = rollmean ( Hardness , k = 9, fill = NA), ma5 = rollmean ( Hardness , k = 11, fill = NA), ma6 = rollmean ( Hardness , k = 13, fill = NA), ma7 = rollmean ( Hardness , k = 15, fill = NA))
```

Next, we visualize the data:

```
ma %>%
gather ( metric , Hardness , Hardness:ma7) %>% ggplot ( aes ( id , Hardness , color = metric )) + geom_line ()
```

```
ma1 = rollmean ( water$Hardness , k = 3) ma2 = rollmean ( water$Hardness , k = 5) ma3 = rollmean ( water$Hardness , k =
7) ma4 = rollmean ( water$Hardness , k = 9) ma5 = rollmean ( water$Hardness , k = 11) ma6 = rollmean ( water$Hardness ,
k = 13) ma7 = rollmean ( water$Hardness , k = 15)
```

Search for turning points:

```
tp1 <- turnpoints (ma1) summary (tp1) tp2 <- turnpoints (ma2) summary (tp2) tp3 <- turnpoints (ma3) summary (tp3) tp4
<- turnpoints (ma4) summary (tp4) tp5 < - turnpoints (ma5) summary (tp5) tp6 <- turnpoints (ma6) summary (tp6) tp7 <-
turnpoints (ma7) summary (tp7)
```

Visualization of turning points for 7 distribution:

```
plot (tp7) plot (ma7, type = "l") lines (tp7)
```

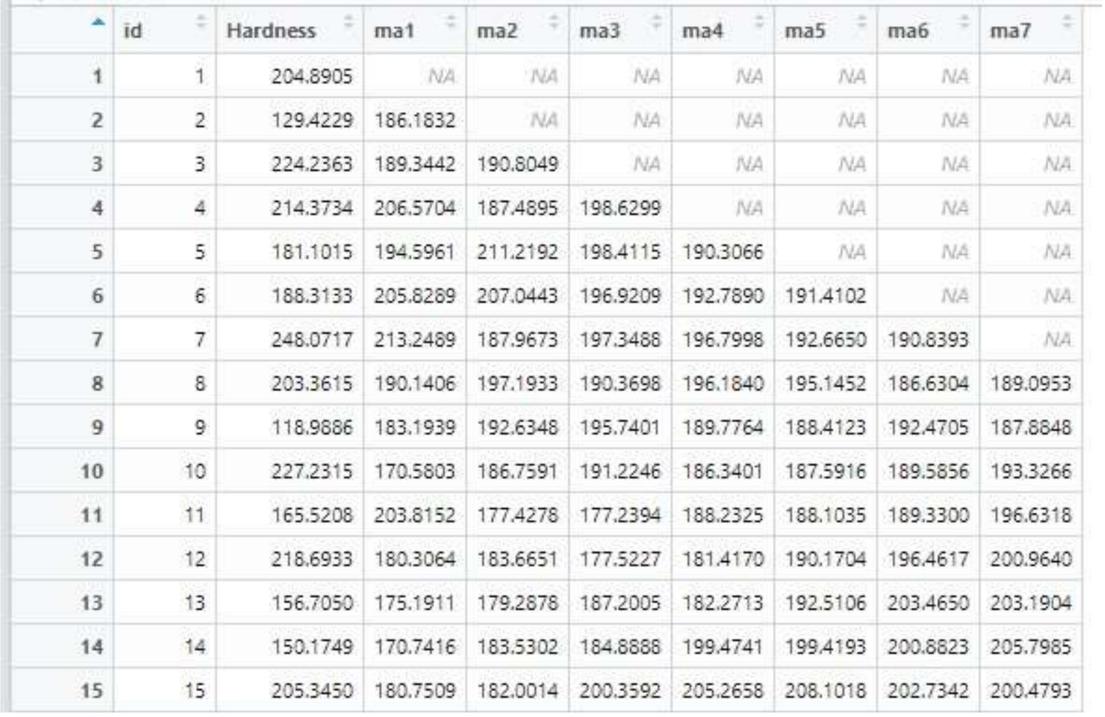
We are looking for correlation coefficients of the smoothed values with the original ones, taking into account that with each smoothing, subtract the columns:

```
cor ( water$Hardness [2:3275],ma1) cor ( water$Hardness [3:3274],ma2) cor ( water$Hardness [4:3273],ma3) cor (
water$Hardness [5:3272], ma4) cor ( water$Hardness [6:3271],ma5) cor ( water$Hardness [7:3270],ma6) cor (
water$Hardness [8:3269],ma7)
```

We smooth the data using the size of the smoothing interval w = 3, then we smooth the obtained smoothed data again, but use the size of the smoothing interval w = 5. Continue the smoothing of the received data with the smoothing interval w = 7 and so on until w = 15. We should get seven columns in a row:

```
maRecursive <- water %>% select ( id , Hardness ) %>% mutate (ma1 = rollmean ( Hardness , k = 3, fill = NA), ma2 =
rollmean (ma1, k = 5, fill = NA), ma3 = rollmean (ma2, k = 7, fill = NA), ma4 = rollmean (ma3, k = 9, fill = NA), ma5 =
rollmean (ma4, k = 11, fill = NA), ma6 = rollmean (ma5, k = 13, fill = NA), ma7 = rollmean (ma6, k = 15, fill = NA))
```

We smooth the data using the sizes of the smoothing interval w = 3, 5, 7, 9, 11, 13, 15 to obtain seven columns. In order to build a moving average. we took as parasetters hardness and id of each water record. You can see 7 columns based on given intervals.



	id	Hardness	ma1	ma2	ma3	ma4	ma5	ma6	ma7
1	1	204.8905	NA	NA	NA	NA	NA	NA	NA
2	2	129.4229	186.1832	NA	NA	NA	NA	NA	NA
3	3	224.2363	189.3442	190.8049	NA	NA	NA	NA	NA
4	4	214.3734	206.5704	187.4895	198.6299	NA	NA	NA	NA
5	5	181.1015	194.5961	211.2192	198.4115	190.3066	NA	NA	NA
6	6	188.3133	205.8289	207.0443	196.9209	192.7890	191.4102	NA	NA
7	7	248.0717	213.2489	187.9673	197.3488	196.7998	192.6650	190.8393	NA
8	8	203.3615	190.1406	197.1933	190.3698	196.1840	195.1452	186.6304	189.0953
9	9	118.9886	183.1939	192.6348	195.7401	189.7764	188.4123	192.4705	187.8848
10	10	227.2315	170.5803	186.7591	191.2246	186.3401	187.5916	189.5856	193.3266
11	11	165.5208	203.8152	177.4278	177.2394	188.2325	188.1035	189.3300	196.6318
12	12	218.6933	180.3064	183.6651	177.5227	181.4170	190.1704	196.4617	200.9640
13	13	156.7050	175.1911	179.2878	187.2005	182.2713	192.5106	203.4650	203.1904
14	14	150.1749	170.7416	183.5302	184.8888	199.4741	199.4193	200.8823	205.7985
15	15	205.3450	180.7509	182.0014	200.3592	205.2658	208.1018	202.7342	200.4793

Fig. 11. Smoothed data according to formulas from Kendel

Visualization of smoothing:

```
maRecursive %>% gather ( metric , Hardness , Hardness:ma7) %>% ggplot ( aes ( id , Hardness , color = metric )) +
geom_line () maR1 = maRecursive$ma1[!is.na(maRecursive$ma1)] maR2 = maRecursive$ma2[!is.na(maRecursive$ma2)]
maR3 = maRecursive$ma3[!is.na(maRecursive$ma3)] maR4 = maRecursive$ma4[!is.na(maRecursive$ma4)] maR5 =
maRecursive$ma5[!is.na(maRecursive$ma5)] maR6 = maRecursive$ma6[!is.na(maRecursive$ma6)] maR7 =
maRecursive$ma7[!is.na(maRecursive$ma7)]
```

Search for turning points:

```
 tpR1 <- turnpoints (maR1) summary (tpR1) tpR2 <- turnpoints (maR2) summary (tpR2) tpR3 <- turnpoints (maR3) summary
(tpR3) tpR4 <- turnpoints (maR4) summary (tpR4) tpR5 < - turnpoints (maR5) summary (tpR5) tpR6 <- turnpoints (maR6)
summary (tpR6) tpR7 <- turnpoints (maR7) summary (tpR7)
```

Visualization of turning points: plot (tpR7) plot (maR7, type = "l") lines (tpR7)

We are looking for correlation coefficients of the smoothed values with the original ones, taking into account that with each smoothing subtract the columns:

cor (water$Hardness [2:3275],maR1) cor (water$Hardness [4:3273],maR2) cor (water$Hardness [7:3270],maR3) cor (water$Hardness [11:3266], maR4) cor (water$Hardness [16:3261],maR5) cor (water$Hardness [22:3255],maR6) cor (water$Hardness [29:3248],maR7)

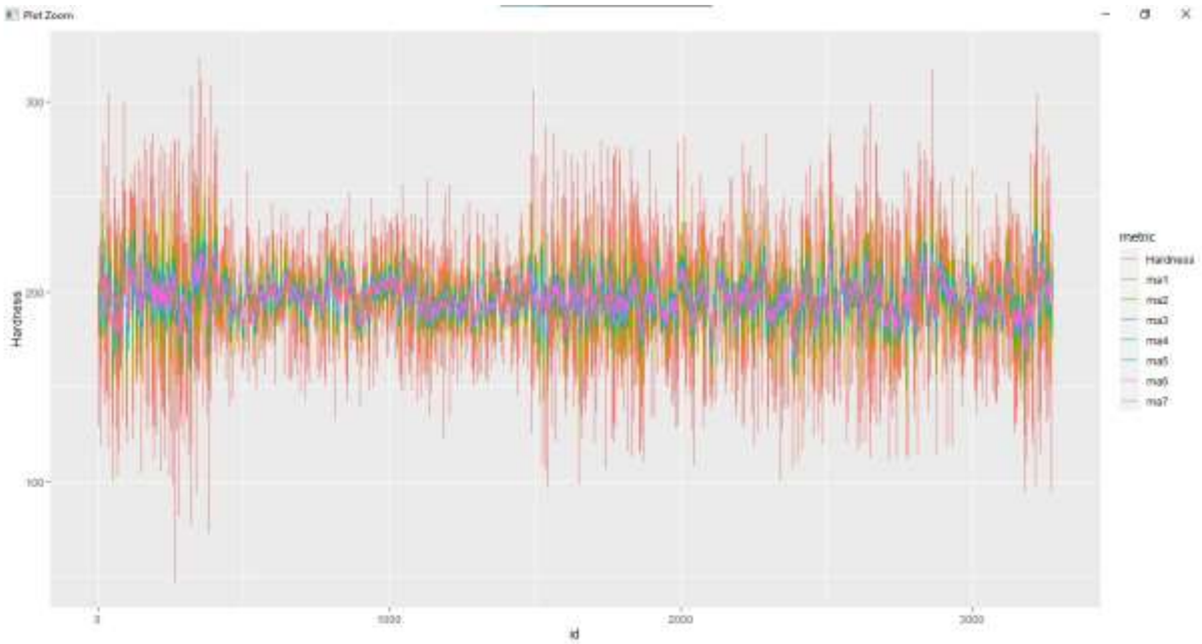


Fig. 12. Graphic representation of smoothed data

From this graph, you can see the hardness parameter fluctuations over the entire interval. The main thing here is hardness and ma7. we see that there is a certain trend here. It's hard to see from the graph, but the end result is a more smooth description of the data.

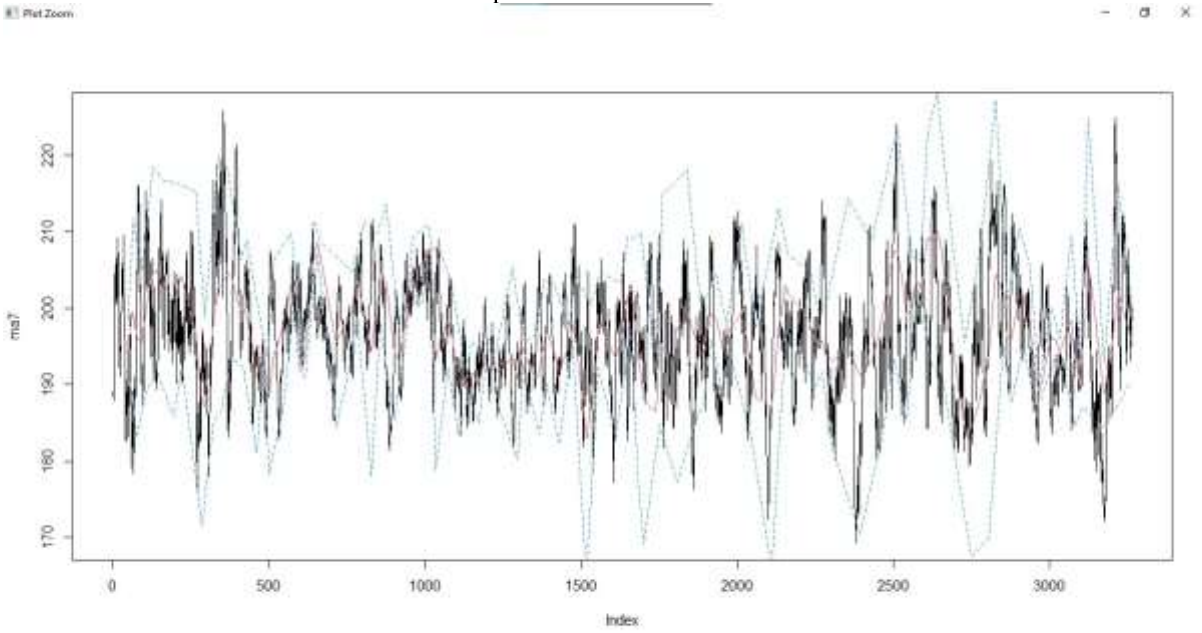
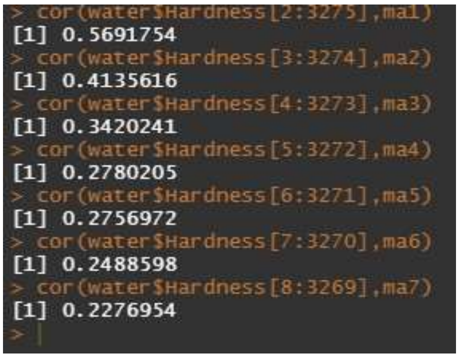


Fig. 13. Visualization of turning points

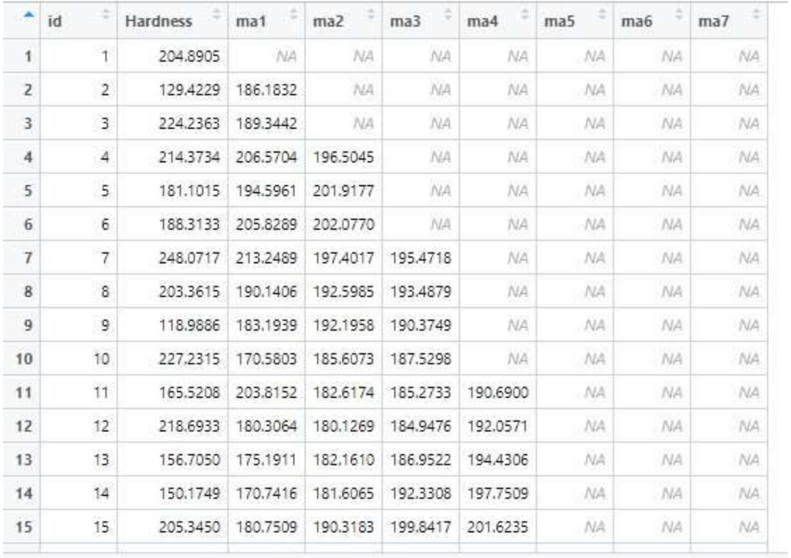
The turning points are quite numerous and detailed smoothing interval increases, the correlation coefficient decreases, because the data is increasingly modified.



```
> cor(water$Hardness[2:3275],ma1)
[1] 0.5691754
> cor(water$Hardness[3:3274],ma2)
[1] 0.4135616
> cor(water$Hardness[4:3273],ma3)
[1] 0.3420241
> cor(water$Hardness[5:3272],ma4)
[1] 0.2780205
> cor(water$Hardness[6:3271],ma5)
[1] 0.2756972
> cor(water$Hardness[7:3270],ma6)
[1] 0.2488598
> cor(water$Hardness[8:3269],ma7)
[1] 0.2276954
> 
```

Fig. 14. Correlation coefficients between smoothed and original data

We smooth the data using the size of the smoothing interval w = 3; then we smooth the obtained smoothed data again using the size of the smoothing interval w = 5. We continue the smoothing of the received data with the smoothing interval *w* = 7 and so on until *w* = 15.



	id	Hardness	ma1	ma2	ma3	ma4	ma5	ma6	ma7
1	1	204.8905	NA	NA	NA	NA	NA	NA	NA
2	2	129.4229	186.1832	NA	NA	NA	NA	NA	NA
3	3	224.2363	189.3442	NA	NA	NA	NA	NA	NA
4	4	214.3734	206.5704	196.5045	NA	NA	NA	NA	NA
5	5	181.1015	194.5961	201.9177	NA	NA	NA	NA	NA
6	6	188.3133	205.8289	202.0770	NA	NA	NA	NA	NA
7	7	248.0717	213.2489	197.4017	195.4718	NA	NA	NA	NA
8	8	203.3615	190.1406	192.5985	193.4879	NA	NA	NA	NA
9	9	118.9886	183.1939	192.1958	190.3749	NA	NA	NA	NA
10	10	227.2315	170.5803	185.6073	187.5298	NA	NA	NA	NA
11	11	165.5208	203.8152	182.6174	185.2733	190.6900	NA	NA	NA
12	12	218.6933	180.3064	180.1269	184.9476	192.0571	NA	NA	NA
13	13	156.7050	175.1911	182.1610	186.9522	194.4306	NA	NA	NA
14	14	150.1749	170.7416	181.6065	192.3308	197.7509	NA	NA	NA
15	15	205.3450	180.7509	190.3183	199.8417	201.6235	NA	NA	NA

Fig. 15. Smoothed data according to formulas from Kendel

It can be seen that we lost more rows and got less accurate data.

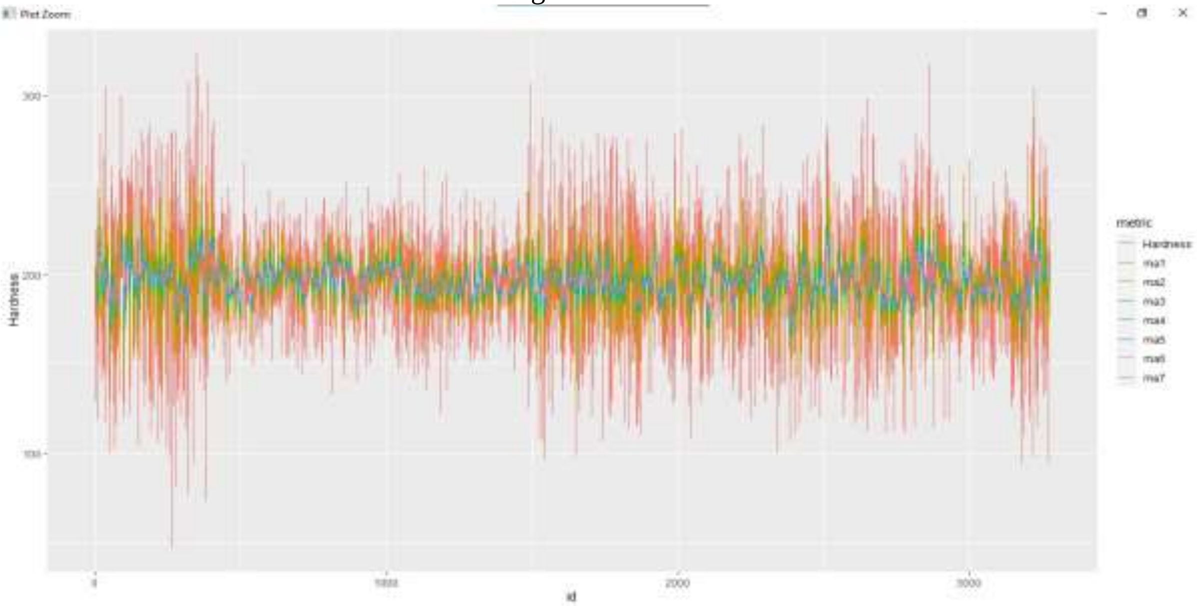
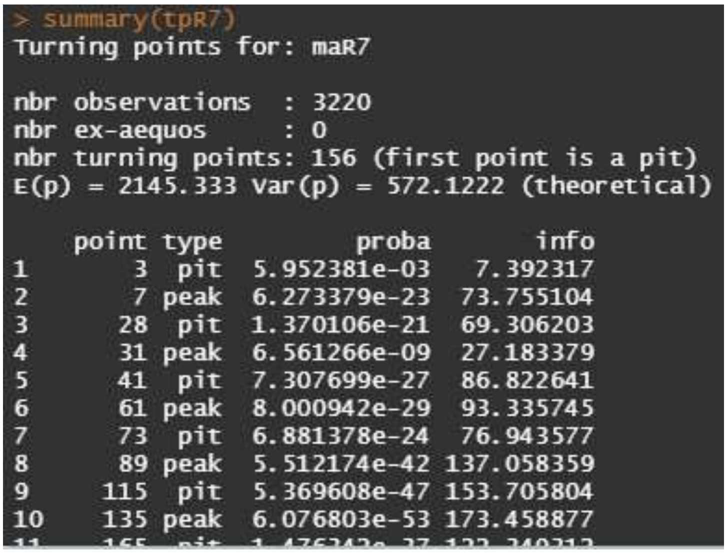


Fig. 16. Graphic representation of smoothed data



```
> summary(tpR7)
Turning points for: maR7

nbr observations  : 3220
nbr ex-aequos     : 0
nbr turning points: 156 (first point is a pit)
E(p) = 2145.333 Var(p) = 572.1222 (theoretical)

    point type        proba       info
1       3  pit  5.952381e-03   7.392317
2       7 peak  6.273379e-23  73.755104
3      28  pit  1.370106e-21  69.306203
4      31 peak  6.561266e-09  27.183379
5      41  pit  7.307699e-27  86.822641
6      61 peak  8.000942e-29  93.335745
7      73  pit  6.881378e-24  76.943577
8      89 peak  5.512174e-42 137.058359
9     115  pit  5.369608e-47 153.705804
10    135 peak  6.076803e-53 173.458877
11    165  pit  1.476343e-37 122.340313
```

Fig. 17. Turning points at the smoothing interval w = 15

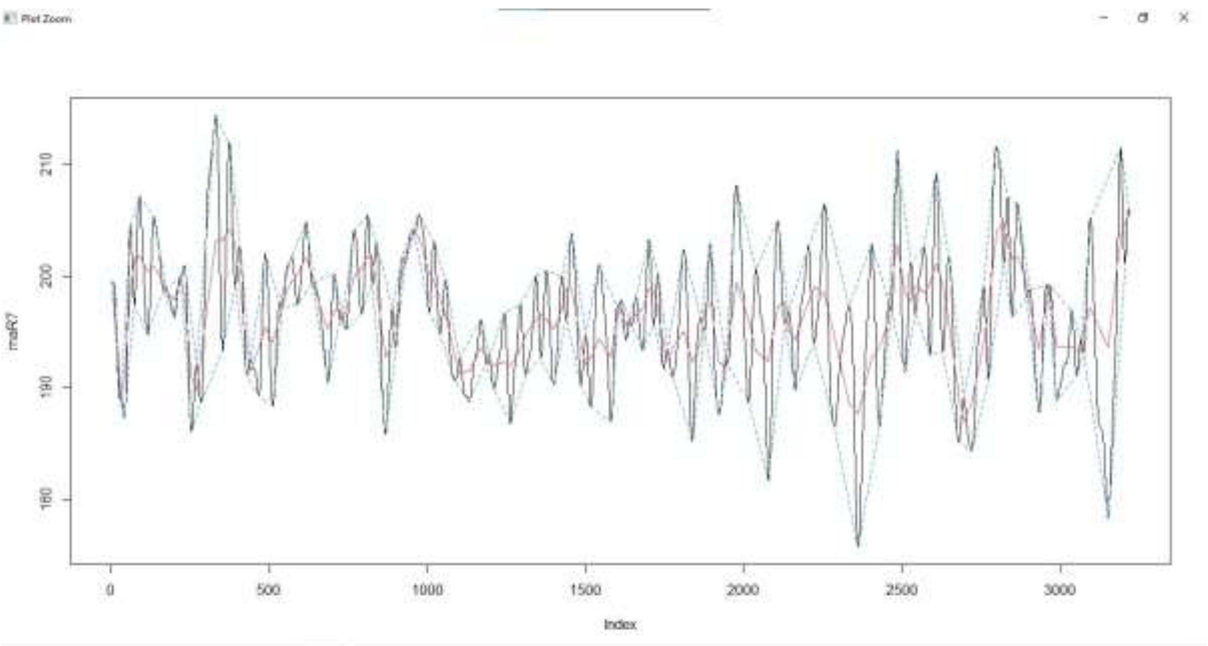
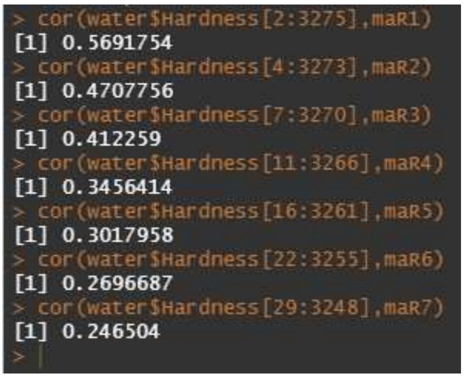


Fig. 18. Visualization of turning points



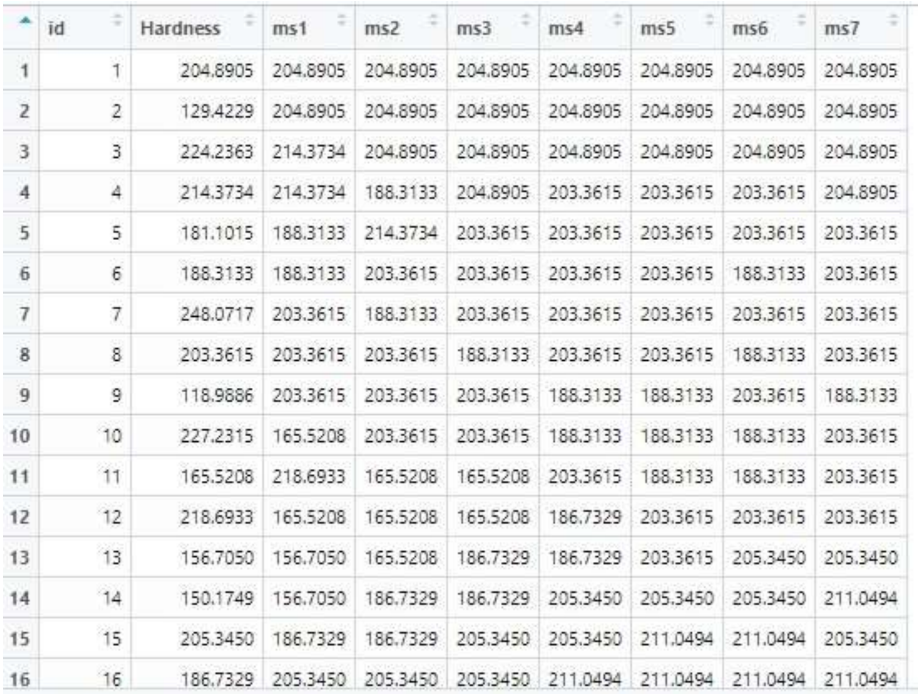
```
> cor(water$Hardness[2:3275],maR1)
[1] 0.5691754
> cor(water$Hardness[4:3273],maR2)
[1] 0.4707756
> cor(water$Hardness[7:3270],maR3)
[1] 0.412259
> cor(water$Hardness[11:3266],maR4)
[1] 0.3456414
> cor(water$Hardness[16:3261],maR5)
[1] 0.3017958
> cor(water$Hardness[22:3255],maR6)
[1] 0.2696687
> cor(water$Hardness[29:3248],maR7)
[1] 0.246504
> 
```

Fig. 19. Correlation coefficients between smoothed and original data

The correlation coefficients also differ, but not much, so the relationship with the raw data remains approximately the same.

2. Median smoothing [67]. The content of the time series's median smoothing algorithm consists of the median's defined values for the smoothing interval levels. Next, the time series level value corresponding to the middle of the smoothing interval is replaced by the median value. Median smoothing completely removes single extreme or anomalous values of levels that are separated from each other by at least half of the smoothing interval; preserves sharp changes in the trend (moving average and exponential smoothing smooth them); effectively removes single levels with very large or very small values that are random and stand out sharply from other levels. We smooth the data using the sizes of the smoothing interval w = 3, 5, 7, 9, 11, 13, 15 to obtain seven columns using the runmed() function:

```
ms <- water %>% select ( id , Hardness ) %>% mutate (ms1 = runmed ( Hardness , 3), ms2 = runmed ( Hardness , 5), ms3 = runmed ( Hardness , 7), ms4 = runmed ( Hardness, 9), ms5 = runmed (Hardness, 11), ms6 = runmed (Hardness, 13), ms7 = runmed (Hardness , 15))
```



	id	Hardness	ms1	ms2	ms3	ms4	ms5	ms6	ms7
1	1	204.8905	204.8905	204.8905	204.8905	204.8905	204.8905	204.8905	204.8905
2	2	129.4229	204.8905	204.8905	204.8905	204.8905	204.8905	204.8905	204.8905
3	3	224.2363	214.3734	204.8905	204.8905	204.8905	204.8905	204.8905	204.8905
4	4	214.3734	214.3734	188.3133	204.8905	203.3615	203.3615	203.3615	204.8905
5	5	181.1015	188.3133	214.3734	203.3615	203.3615	203.3615	203.3615	203.3615
6	6	188.3133	188.3133	203.3615	203.3615	203.3615	203.3615	188.3133	203.3615
7	7	248.0717	203.3615	188.3133	203.3615	203.3615	203.3615	203.3615	203.3615
8	8	203.3615	203.3615	203.3615	188.3133	203.3615	203.3615	188.3133	203.3615
9	9	118.9886	203.3615	203.3615	203.3615	188.3133	188.3133	203.3615	188.3133
10	10	227.2315	165.5208	203.3615	203.3615	188.3133	188.3133	188.3133	203.3615
11	11	165.5208	218.6933	165.5208	165.5208	203.3615	188.3133	188.3133	203.3615
12	12	218.6933	165.5208	165.5208	165.5208	186.7329	203.3615	203.3615	203.3615
13	13	156.7050	156.7050	165.5208	186.7329	186.7329	203.3615	205.3450	205.3450
14	14	150.1749	156.7050	186.7329	186.7329	205.3450	205.3450	205.3450	211.0494
15	15	205.3450	186.7329	186.7329	205.3450	205.3450	211.0494	211.0494	205.3450
16	16	186.7329	205.3450	205.3450	205.3450	211.0494	211.0494	211.0494	211.0494

Fig. 20. Median smoothed data

We used the same smoothing intervals and operations as in the previous point.

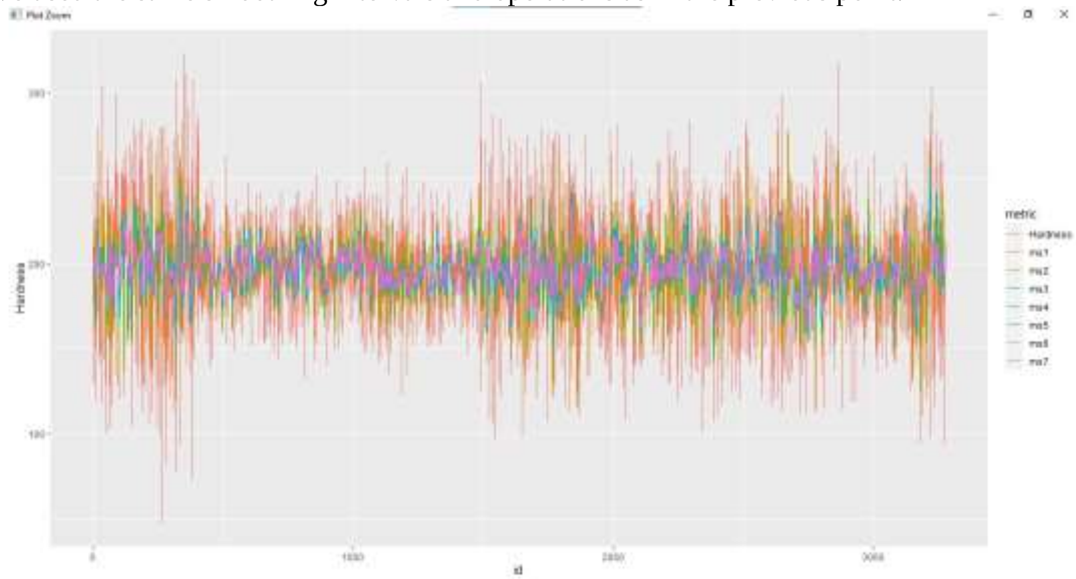


Fig. 21. Graphic representation of smoothed data

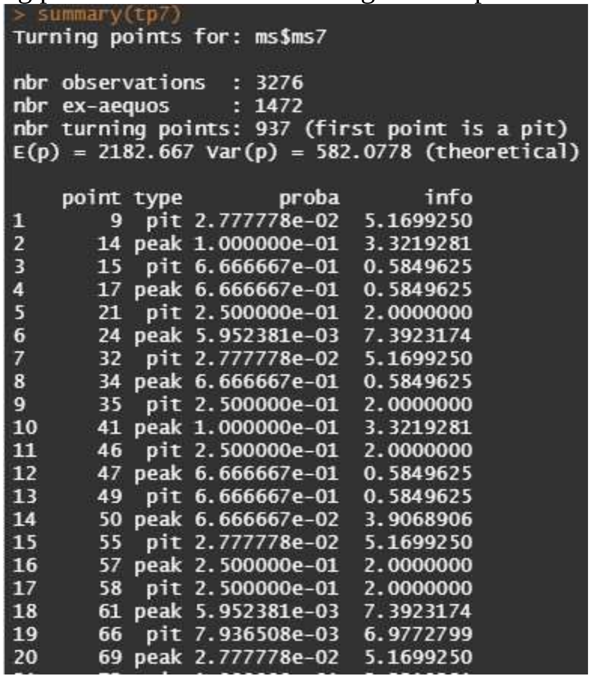
Visualization of smoothing:

ms %>%
gather (metric , Hardness , Hardness:ms7) %>% ggplot (aes (id , Hardness , color = metric)) + geom_line () Turnpoints search: tp1 <- turnpoints (ms$ms1) summary (tp1) tp2 <- turnpoints (ms$ms2) summary (tp2) tp3 <- turnpoints (ms$ms3) summary (tp3) tp4 <- turnpoints (ms$ms4) summary (tp4) tp5 <- turnpoints (ms$ms5) summary (tp5)
tp6 <- turnpoints (ms$ms6) summary (tp6) tp7 <- turnpoints (ms$ms7) summary (tp7)

Visualization of turning points:

plot (tp7) plot (ms$ms7, type = "l") lines (tp7)

Now let's find the turning points for the last smoothing with step 15:



```
> summary(tp7)
Turning points for: ms$ms7

nbr observations  : 3276
nbr ex-aequos     : 1472
nbr turning points: 937 (first point is a pit)
E(p) = 2182.667 Var(p) = 582.0778 (theoretical)

    point type        proba      info
1       9  pit 2.777778e-02  5.1699250
2      14 peak 1.000000e-01  3.3219281
3      15  pit 6.666667e-01  0.5849625
4      17 peak 6.666667e-01  0.5849625
5      21  pit 2.500000e-01  2.0000000
6      24 peak 5.952381e-03  7.3923174
7      32  pit 2.777778e-02  5.1699250
8      34 peak 6.666667e-01  0.5849625
9      35  pit 2.500000e-01  2.0000000
10     41 peak 1.000000e-01  3.3219281
11     46  pit 2.500000e-01  2.0000000
12     47 peak 6.666667e-01  0.5849625
13     49  pit 6.666667e-01  0.5849625
14     50 peak 6.666667e-02  3.9068906
15     55  pit 2.777778e-02  5.1699250
16     57 peak 2.500000e-01  2.0000000
17     58  pit 2.500000e-01  2.0000000
18     61 peak 5.952381e-03  7.3923174
19     66  pit 7.936508e-03  6.9772799
20     69 peak 2.777778e-02  5.1699250
```

Fig. 21. Turning points at the smoothing interval w = 15

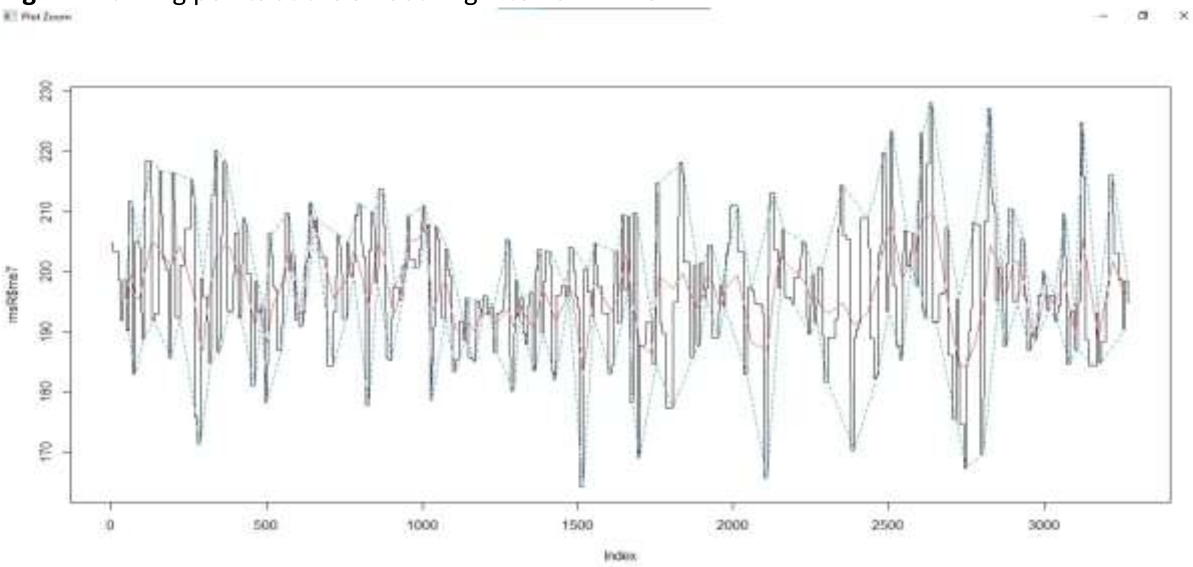


Fig. 22. Visualization of turning points

Correlation coefficients of smoothed values with original ones:

cor (water$Hardness,ms$ms1) cor (water$Hardness,ms$ms2) cor (water$Hardness,ms$ms3) cor (water$Hardness,ms$ms4) cor (water$Hardness,ms$ms5) cor (water$Hardness,ms$ms6) cor (water$Hardness,ms$ms7)

We smooth the data using the size of the smoothing interval w = 3, then we smooth the obtained smoothed data again, but use the size of the smoothing interval w = 5. Continue the smoothing of the received data with the smoothing interval w = 7 and so on until w = 15. We should get seven columns in a row:

```
msR <- water %>% select ( id , Hardness ) %>% mutate (ms1 = runmed ( Hardness , 3), ms2 = runmed (ms1, 5), ms3 =
runmed (ms2, 7), ms4 = runmed (ms3 , 9), ms5 = runmed (ms4, 11), ms6 = runmed (ms5, 13), ms7 = runmed (ms6, 15))
```

Visualization of smoothing:

```
msR %>%
gather ( metric , Hardness , Hardness:ms7) %>% ggplot ( aes ( id , Hardness , color = metric )) + geom_line () Turnpoints
search: tp1 <- turnpoints (msR$ms1) summary (tp1) tp2 <- turnpoints (msR$ms2) summary (tp2) tp3 <- turnpoints
(msR$ms3) summary (tp3)
tp4 <- turnpoints (msR$ms4) summary (tp4) tp5 <- turnpoints (msR$ms5) summary (tp5) tp6 <- turnpoints (msR$ms6)
summary (tp6) tp7 <- turnpoints (msR$ms7) summary ( tp7)
```

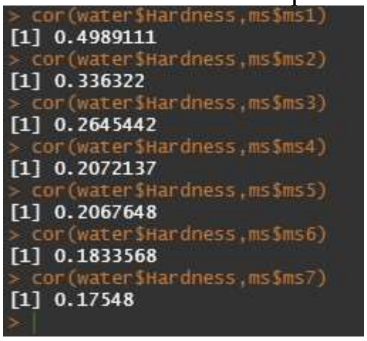
Visualization of turning points:

```
plot (tp7) plot (msR$ms7, type = "l") lines (tp7)
```

Correlation coefficients of smoothed values with original ones:

```
cor (water$Hardness,msR$ms1) cor (water$Hardness,msR$ms2) cor (water$Hardness,msR$ms3) cor
(water$Hardness,msR$ms4) cor (water$Hardness,msR$ms5) cor (water$Hardness,msR$ms6) cor
(water$Hardness,msR$ms7)
```

The graph looks exactly like this because the data has acquired a complete form.



```
> cor(water$Hardness,ms$ms1)
[1] 0.4989111
> cor(water$Hardness,ms$ms2)
[1] 0.336322
> cor(water$Hardness,ms$ms3)
[1] 0.2645442
> cor(water$Hardness,ms$ms4)
[1] 0.2072137
> cor(water$Hardness,ms$ms5)
[1] 0.2067648
> cor(water$Hardness,ms$ms6)
[1] 0.1833568
> cor(water$Hardness,ms$ms7)
[1] 0.17548
>
```

Fig. 23. Correlation coefficients between smoothed and original data

The correlation coefficient is smaller than the data of the previous methods, which means that this method is not quite suitable for the given dataset because it reduces its reliability.

Correlation analysis [66-80] is a group of methods that allow detecting the presence and degree of relationship between several randomly changing parameters. Special numerical characteristics and their statistics assess the degree of such a relationship. The correlation appears in the form of a tendency to change the average values of the function depending on changes in the argument. ggpubr library - it is a library for data visualization in R. We build a correlation field:

```
library ( ggpubr )
plot ( water$ph , water$Solids , main =" Correlation field ", xlab =" Age ",
ylab = " Cholesterol ")
```

From the graphically presented field, it can be concluded that the indicators correlate quite strongly [55].

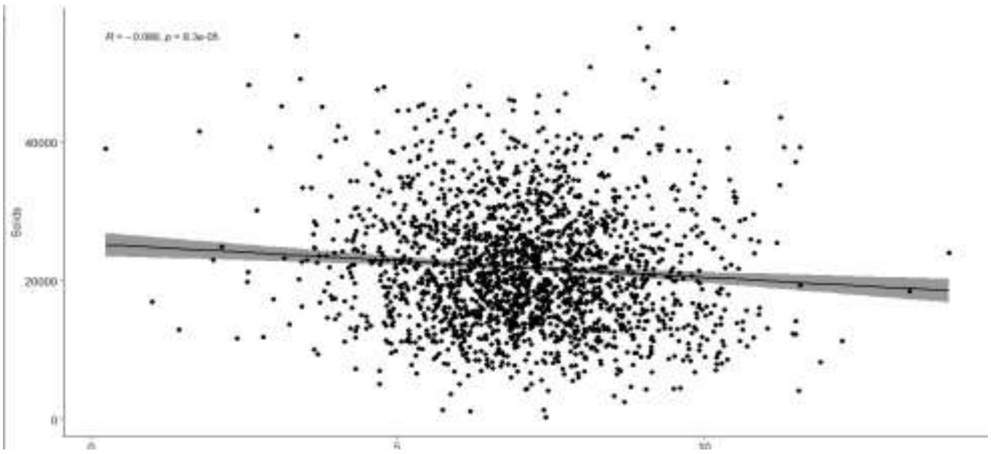


Fig. 24. Correlation field

We determine the correlation coefficient:

```
correlation <- cor ( water$ph , water$Solids )
```

Using the ggscatter method of the ggrubr library , we calculate correlation relation:

```
qwe <- ggscatter ( water , x = " ph ", y = " Solids ", add = " reg.line ", conf.int = TRUE, cor.coef = TRUE, cor.method = " person ", xlab = " ph ", ylab = " Solids ")
```

We divide the data into 3 parts:

```
ph1 <- water$ph [1:1092] ph2 <- water$ph [1093:2184] ph3 <- water$ph [2185:3276]
```

For parts, we build a correlation matrix (rcorr): mydata.rcorr = rcorr (as.matrix (cbind (ph1, ph2, ph3)))

We find multiple correlation coefficients:

```
numericData <- cbind ( water$id,water$ph , water$Hardness , water$Solids ,    water$Chloramines , water$Sulfate,water$Conductivity,water$Organic_carbon,water$Trihalome thanes,water$Turbidity ) chart.Correlation ( numericData , histogram =TRUE, pch =19)
```

Let's plot graphs of autocorrelation functions using acf :

```
data <- cbind ( water$ph , water$Solids ) colnames ( data ) <- c(" ph ", " Solids ")
autocorrelation <- acf ( data , lag.max = 1, type = c(" correlation "),
plot = TRUE, xlab =" ph ", ylab =" Solids ")
```

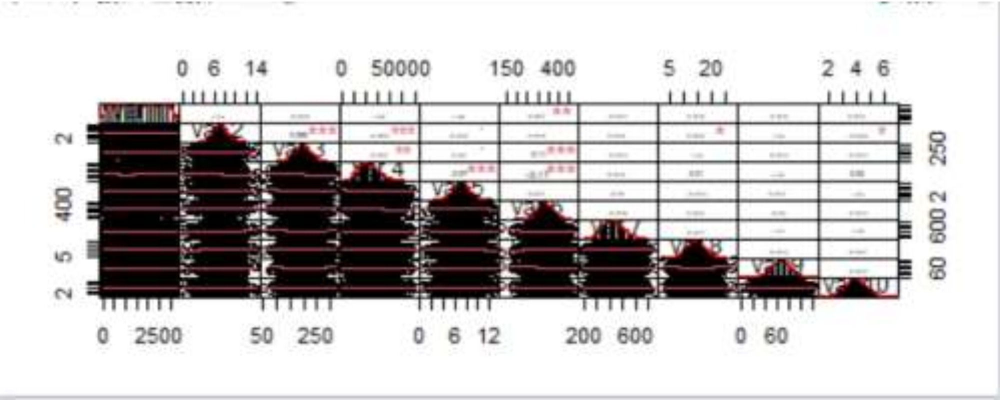


Fig. 25. Correlation matrix

The matrix displays all the coefficients and even graphically displays the relationships. Multiple correlation coefficients show that the dataset has weak but present relationships, based on which results can be constructed. Cluster analysis is one of the methods of multivariate statistical analysis; that is, each observation is represented not by a single indicator but by a set of values of various indicators [5, 86, 91-99]. It includes algorithms with the help of which the clusters' formation and the distribution of objects by clusters are carried out. Cluster analysis, first of all, solves the problem of adding structure to the data and also ensures the selection of groups of objects, that is, looks for the division of the population into areas of accumulation of objects. Cluster analysis allows you to consider fairly significant volumes of data, sharply shorten and compress them, make them compact.

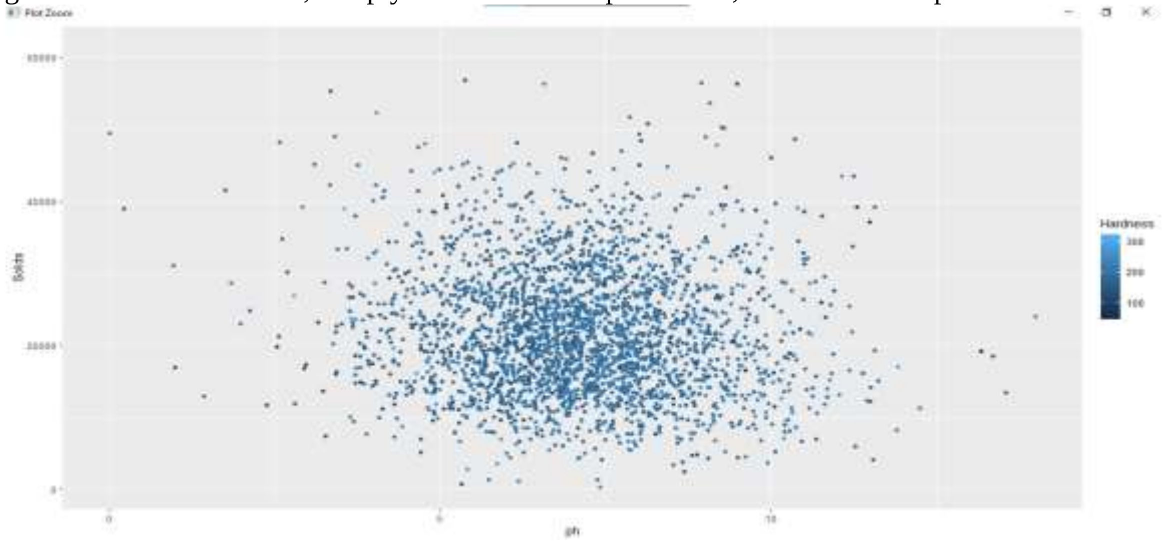


Fig. 26. Graphic representation of cluster analysis

Because we use the RStudio environment and the R language to perform the laboratory work in order to build clusters, it is not necessary to form an "object-property" table from the provided data, to form from the closely located "original table" and "table-copy", to build a proximity matrix and the like. We can immediately perform the cluster analysis procedure.

Performing a cluster analysis procedure using built-in R methods:

Let's select the parameters MaxHR, Cholesterol and ChestPainType and build a graphical representation of the clustering: factoextra - The library provides some easy-to-use functions to extract and visualize the results of multivariate data analysis.

```
library (ggplot2) library ( factoextra ) library ( rEMM ) ggplot ( water , aes ( ph , Solids , col = Hardness )) + geom_point ()
Let's build the clustering matrix: set.seed (55) cluster <- kmeans ( cbind ( water$ ph , water$Solids ), 3, nstart = 10) cluster
table ( cluster$cluster,water$Hardness )
build a dendrogram : data <- cbind ( water$ph , water$Solids )
        data.hclust =hclust(dist(scale(data,center=apply(data,2,mean),scale=apply(data,2,sd)))) plot ( data.hclust )
```

We chose the parameters Solids, Hardness and built a graphical representation of the clustering:

6. Conclusions

The work establishes the main trends in determining the suitability of water for human consumption: the most common indicator of the acid-base balance of water is from 6 to 7, most of our data set are not suitable for drinking water, the most common indicator of the sulfate balance of water is from 300 to 350, the most common indicators of the carbon balance of water are within 12-15. The average and most popular value of the acid-alkaline balance of water is 7; the standard deviation from this parameter is insignificant, the indicators vary in the range of 0-14, and the sign of the acid-alkaline balance of water is quite stable. In this work, we constructed graphs in Cartesian and polar coordinate systems, derived quantitative characteristics of descriptive statistics, and formed histograms and cumulates. Investigating this problem, we used the main methods of visualization, graphic representation and primary statistical processing of numerical data. Methods of correlation analysis of experimental data presented by time sequences were also used in work.

The most common indicator values determined by histograms:

- The most common indicator of the acid-alkaline balance of water is from 6 to 7;
- Most of our data set are non-potable water;
- The most common indicator of the sulfate balance of water is from 300 to 350;
- The most common indicators of the carbon balance of water are in the range of 12-15.

As can be seen from the histogram in Fig. 33, most of the studied water from our dataset is unsuitable for consumption (more than 1200 records).

The results of the descriptive statistics of the level of acidity are the following data:

- Average is 7.08599;
- The standard error is 0.035;
- Median is 7.027297;
- Fashion is 8.316766;
- The standard deviation is 1.573337;
- Sample variance is 2.474157;
- Skewness is 0.6185764;
- Asymmetry is 0.04891027;
- Interval is 13.7725;
- The minimum is 0.23;
- The maximum is 14;
- The amount is 14249.93;
- Volume (quantity) is 2011;
- Coefficient of variation is 22.2%.

After finding some statistical data for the water acidity level, we saw that this level ranges from 5 to 9. The level of acidity should be in the range of 6.5 - 8.5. We see an average value of 7, which is within these limits; the standard error is relatively small. The median also falls within these limits.

We see a minimum of 0.23, which is completely abnormal and can almost be equated to car battery acid, and a maximum of 14, which can be equated to soapy water. The difference between the maximum and the minimum is the indicator - the interval, which in our case is 13.7725. Consider the indicator - kurtosis. For a normal distribution, the kurtosis is zero. If the kurtosis of some distribution is different from zero, then this distribution's density curve differs from the normal distribution's density curve. Since our kurtosis is positive, the theoretical curve has a higher and "sharper" peak than the normal curve. Otherwise, this curve would have a theoretically lower and flatter peak than the normal curve.

The value of the variation parameter can provide interesting information - this is the difference in the numerical values of the characteristics of the population units and their fluctuations around the average value that characterizes the population. The smaller the variation, the more homogeneous the population and the more reliable (typical) the average value. If the variation percentage is lower than 33%, then the data set is quantitatively homogeneous, which corresponds to our result of 22.2%. You can also form certain facts based on our results:

- The average and most popular value of the acid-alkaline balance of water is 7;
- The standard deviation from this parameter is insignificant;
- Indicators range from 0.23 to 14;
- The sign that the acid-alkaline balance of water is quite stable.

7. References

[1] O. Kuzmin, M. Bublyk, A. Shakhno, O. Korolenko, H. Lashkun, Innovative development of human capital in the conditions of globalization, E3S Web of Conferences 166 (2020) 13011.

[2] O. Ilyash, O. Yildirim, D. Doroshkevych, L. Smoliar, T. Vasyltsiv, R. Lupak, Evaluation of enterprise investment attractiveness under circumstances of economic development, Bulletin of Geography. Socio-economic Series 47 (2020) 95-113. http://doi.org/10.2478/bog-2020-0006.

[3] I. Jonek-Kowalska, Housing Infrastructure as a Determinant of Quality of Life in Selected Polish Smart Cities Smart Cities 5(3) (2022) 924–946.

[4] O. Maslak, V. Danylko, M. Skliar, Automation and Digitalization of Quality Cost Management of Power Engineering Enterprises, in: Proceedings of the 25th IEEE International Conference on Problems of Automated Electric Drive. Theory and Practice, PAEP 2020. https://doi.org/10.1109/MEES52427.2021.9598744

[5] M. Bublyk, A. Kowalska-Styczen, V. Lytvyn, V. Vysotska, The Ukrainian Economy Transformation into the Circular Based on Fuzzy-Logic Cluster Analysis, Energies 14 (2021) 5951. doi: https://doi.org/10.3390/en14185951.

[6] R. Yurynets, Z. Yurynets, O. Budiakova, L. Gnylianska, M. Kokhan, Innovation and Investment Factors in the State Strategic Management of Social and Economic Development of the Country. Modeling and Forecasting, CEUR Workshop Proceedings Vol-2917 (2021) 357-372.

[7] Y. Matseliukh, V. Vysotska, M. Bublyk, T. Kopach, O. Korolenko, Network modelling of resource consumption intensities in human capital management in digital business enterprises by the critical path method, CEUR Workshop Proceedings Vol-2851 (2021) 366–380.

[8] I. Jonek-Kowalska, Towards the Reduction of CO2 Emissions. Paths of Pro-Ecological Transformation of Energy Mixes in European Countries with an Above-Average Share of Coal in Energy Consumption. Resources Policy 77 (2022). doi: 10.1016/j.resourpol.2022.102701.

[9] M. Bublyk, V. Vysotska, Y. Matseliukh, V. Mayik, M. Nashkerska, Assessing losses of human capital due to man-made pollution caused by emergencies, CEUR Workshop Proceedings Vol-2805 (2020) 74-86.

[10] M. Bublyk, Y. Matseliukh, Small-batteries utilization analysis based on mathematical statistics methods in challenges of circular economy, CEUR workshop proceedings Vol-2870 (2021) 1594-1603.

[11] I. Jonek-Kowalska, R. Wolniak, Economic opportunities for creating smart cities in Poland. Does wealth matter?, Cities114 (2021) 103222.

[12] I. Rishnyak, O. Veres, V. Lytvyn, M. Bublyk, I. Karpov, V. Vysotska, V. Panasyuk, Implementation models application for IT project risk management, CEUR Workshop Proceedings Vol-2805 (2020) 102-117.

[13] O. Kuzmin, M. Bublyk, Economic evaluation and government regulation of technogenic (Man-Made) damage in the national economy, in: International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2016, pp. 37–39.

[14] R. Wolniak, I. Jonek-Kowalska, The level of the quality of life in the city and its monitoring Innovation, The European Journal of Social Science Research 34(3) (2021) 376–398.

[15] T. Vasyltsiv, I. Irtyshcheva, R. Lupak, N. Popadynets, Y. Shyshkova, Y. Boiko, O. Ishchenko, Economy's innovative technological competitiveness: Decomposition, methodic of analysis and priorities of public policy, Management Science Letters 10(13) (2020) 3173-3182. https://doi.org/10.5267/j.msl.2020.5.004.

[16] M. S. I. Khan, N. Islam, J. Uddin, S. Islam, M. K. Nasir, Water quality prediction and classification based on principal component regression and gradient boosting classifier approach, Journal of King Saud University-Computer and Information Sciences 34(8) (2022) 4773-4781. https://doi.org/10.1016/j.jksuci.2021.06.003.

[17] T. H. H Aldhyani, M. Al-Yaari, H. Alkahtani, M. Maashi, Water Quality Prediction Using Artificial Intelligence Algorithms, Applied Bionics and Biomechanics 2020, Article ID 6659314, 12 p., 2020. https://doi.org/10.1155/2020/6659314.

[18] S. Chatterjee, S. Sarkar, N. Dey, S. Sen, T. Goto, N. C. Debnath, Water quality prediction: Multi objective genetic algorithm coupled artificial neural network based approach, in: Int. Conf. on Industrial Informatics, 2017, pp. 963-968. https://ieeexplore.ieee.org/document/8104902.

[19] Water quality describes the condition of the water, including chemical, physical, and biological characteristics, usually with respect to its suitability for a particular purpose such as drinking or swimming. URL: https://floridakeys.noaa.gov/ocean/waterquality.html.

[20] Importance of Water Quality and Testing. URL: https://www.cdc.gov/healthywater/drinking/public/water_quality.html.

[21] A. N. Ahmed, et al. Machine learning methods for better water quality prediction, Journal of Hydrology 578 (2019) 124084.

[22] Y. Chen, L. Song, Y. Liu, L. Yang, D. Li, A review of the artificial neural network models for water quality prediction. Applied Sciences 10(17) (2020) 5776.

[23] A. Gozhyj, I. Kalinina, V. Gozhyj, Fuzzy cognitive analysis and modeling of water quality, in: International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2017, pp. 289-294.

[24] M. Linan, B. Gerardo, R. Medina, Self-Organizing Map with Nguyen-Widrow Initialization Algorithm for Groundwater Vulnerability Assessment, International Journal of Computing 19(1) (2020) 63-69.

[25] D.K. Mozgovoy, V.V. Hnatushenko, V.V. Vasyliev, Automated recognition of vegetation and water bodies on theterritory of megacities in satellite images of visible and IR bands, ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. IV-3 (2018) 167–172, https://doi.org/10.5194/isprs-annals-IV-3-167-2018.

[26] W. Wójcik, et. al., Hydroecological investigations of water objects located on urban areas, in: Environmental Engineering V – Proceedings of the 5th National Congress of Environmental Engineering, 2017, pp. 155–160.

[27] R.Ya. Kosarevich, et. al., Assessment of damages caused by thermal fatigue cracks in water economizer collector, Fiziko-Khimicheskaya Mekhanika Materialov 40(1) (2004) 109–115.

[28] O. Alokhina, et. al., Solar Activity and Water Content of Closed Lake Ecosystems, in: General Assembly and Scientific Symposium of the International Union of Radio Science, 2020, 9232274.

[29] N. Anufrieva, Y. Obukh, B. Rusyn, I. Fartushok, Expert computer system for technical diagnostics of the efficiency of main constitutive elements of the water steam route, in: The Experience of Designing and Application of CAD Systems in Microelectronics - Proceedings of the 9th International Conference, CADSM, 2007, pp. 206.

[30] N. Anufrieva, Y. Obukh, B. Rusyn, I. Fartushok, Typical damage image database of the main constitutive elements of the water steam route, in: The Experience of Designing and Application of CAD Systems in Microelectronics - Proceedings of the International Conference, 2007, pp. 518.

[31] R Elmahdi. Predicting Water Quality Variables. URL: https://scholar.sun.ac.za/bitstream/handle/10019.1/108072/elmahdi_predicting_2020.pdf?sequence=2&isAllowed=y.

[32] V. Sagan, K. T. Peterson, M. Maimaitijiang, P. Sidike, J. Sloan, B. A Greeling, S. Maalouf, C. Adams, Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. URL: https://www.sciencedirect.com/science/article/abs/pii/S0012825220302336.

[33] T. S. Kapalanga, Z. Hoko, W. Gumindoga, L. Chikwiramakomo, Remote-sensing-based algorithms for water quality monitoring in Olushandja Dam, north-central Namibia, Water Supply 21(5) (2021) 1878-1894.

[34] Y. F. Zhang, P. J. Thorburn, M. P. Vilas, P. Fitch, Machine learning approaches to improve and predict water quality data, in: International Congress on Modelling and Simulation-Supporting Evidence-Based Decision Making: the Role of Modelling and Simulation, MODSIM 2019.

[35] J. O., Oladipo, A. S., Akinwumiju, O. S., Aboyeji, A. A. Adelodun, Comparison between fuzzy logic and water quality index methods: A case of water quality assessment in Ikare community, Southwestern Nigeria, Environmental Challenges 3 (2021) 100038.

[36] O. S. Aboyeji, S. F. Eigbokhan, Evaluations of groundwater contamination by leachates around Olusosun open dumpsite in Lagos metropolis, southwest Nigeria, Journal of environmental management 183 (2016) 333-341.

[37] M. Yilma, Z. Kiflie, A. Windsperger, N. Gessese, Application of artificial neural network in water quality index prediction: a case study in Little Akaki River, Addis Ababa, Ethiopia, Modeling Earth Systems and Environment 4(1) (2018) 175-187.

[38] D. M. Bushero, Z. A. Angello, B. M. Behailu, Evaluation of hydrochemistry and identification of pollution hotspots of little Akaki river using integrated water quality index and GIS, Environmental Challenges 8 (2022) 100587.

[39] M. F. M. Nasir, M. S. Samsudin, I. Mohamad, M. R. A. Awaluddin, M. A. Mansor, H. Juahir, N. Ramli, River water quality modeling using combined principle component analysis (PCA) and multiple linear regressions (MLR): a case study at Klang River, Malaysia, World Applied Sciences Journal 14 (2011) 73-82.

[40] M. Hameed, S. S. Sharqi, Z. M. Yaseen, H. A. Afan, A. Hussain, A. Elshafie, Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia, Neural Computing and Applications 28(1) (2017) 893-905.

[41] J. Y. Ho, et. al., Towards a time and cost effective approach to water quality index class prediction, Journal of Hydrology 575 (2019) 148-165.

[42] R. Barzegar, M. T. Aalami, J. Adamowski, Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model, Stochastic Environmental Research and Risk Assessment 34(2) (2020) 415-433.

[43] Z. Li, F. Peng, B. Niu, G. Li, J. Wu, Z. Miao, Water quality prediction model combining sparse auto-encoder and LSTM network, in: IFAC-PapersOnLine 51(17) (2018) 831-836.

[44] S. B. H. S. Asadollah, A. Sharafati, D. Motta, Z. M. Yaseen, River water quality index prediction and uncertainty analysis: A comparative study of machine learning models, Journal of environmental chemical engineering 9(1) (2021) 104599.

[45] T. Rajaee, S. Khani, M. Ravansalar, Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: A review, Chemometrics and Intelligent Laboratory Systems 200 (2020) 103978.

[46] M. S. Samsudin, A. Azid, S. I. Khalit, M. S. A. Sani, F. Lananan, Comparison of prediction model using spatial discriminant analysis for marine water quality index in mangrove estuarine zones, Marine pollution bulletin 141 (2019) 472-481.

[47] M. Imani, M. M. Hasan, L. F. Bittencourt, K. McClymont, Z. Kapelan, A novel machine learning application: Water quality resilience prediction Model, Science of the Total Environment 768 (2021) 144459.

[48] M. Ranjithkumar, L. Robert, Machine Learning Techniques and Cloud Computing to Estimate River Water Quality-Survey, Inventive communication and computational technologies, Springer, Singapore, 2021, p. 387-396.

[49] Y. Trach, R. Trach, M. Kalenik, E. Koda, A. Podlasek, A Study of Dispersed, Thermally Activated Limestone from Ukraine for the Safe Liming of Water Using ANN Models, Energies 14(24) (2021) 8377.

[50] Y. Trach, D. Chernyshev, O. Biedunkova, V. Moshynskyi, R. Trach, I. Statnyk, Modeling of Water Quality in West Ukrainian Rivers Based on Fluctuating Asymmetry of the Fish Population, Water 14(21) (2022) 3511.

[51] L. V. Hryhorenko, Drinking water quality influence to the peasants' morbidity in the Ukrainian settlements, International Journal of Statistical Distributions and Applications 3(3) (2017) 38-46.

[52] J. Ober, J. Karwot, S. Rusakov, Tap Water Quality and Habits of Its Use: A Comparative Analysis in Poland and Ukraine, Energies 15(3) (2022) 981.

[53] B. Polishchuk, A. Berko, L. Chyrun, M. Bublyk, V. Schuchmann, The Rain Prediction in Australia Based Big Data Analysis and Machine Learning Technology, in: International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2021, pp. 97–100.

[54] D. Koshtura, M. Bublyk, Y. Matseliukh, D. Dosyn, L. Chyrun, O. Lozynska, I. Karpov, I. Peleshchak, M. Maslak, O. Sachenko, Analysis of the demand for bicycle use in a smart city based on machine learning, CEUR workshop proceedings Vol-2631 (2020) 172-183.

[55] A. Katrenko, I. Krislata, O. Veres, O. Oborska, T. Basyuk, A. Vasyliuk, I. Rishnyak, N. Demyanovskyi, O. Meh, Development of traffic flows and smart parking system for smart city. CEUR Workshop Proceedings Vol-2604 (2020) 730–745.

[56] V. V. Lytvyn, M. I. Bublyk, V. A. Vysotska, Y. R. Matseliukh, Technology of visual simulation of passenger flows in the field of public transport Smart City, Radioelectronics, informatics, management, No. 4, 2021.

[57] L. Podlesna, M. Bublyk, I. Grybyk, Y. Matseliukh, Y. Burov, P. Kravets, O. Lozynska, I. Karpov, I. Peleshchak, R. Peleshchak, Optimization model of the buses number on the route based on queueing theory in a Smart City, CEUR Workshop Proceedings Vol-2631 (2020) 502-515.

[58] Y. Matseliukh, M. Bublyk, V. Vysotska, Development of intelligent system for visual passenger flows simulation of public transport in smart city based on neural network, CEUR Workshop Proceedings, Vol-2870 (2021).

[59] V. Husak, L. Chyrun, Y. Matseliukh, A. Gozhyj, R. Nanivskyi, M. Luchko, Intelligent Real-Time Vehicle Tracking Information System, CEUR Workshop Proceedings 2917 (2021) 666-698.

[60] V. Lytvyn, A. Hryhorovych, V. Hryhorovych, L. Chyrun, V. Vysotska, M. Bublyk, Medical Content Processing in Intelligent System of District Therapist, CEUR Workshop Proceedings Vol-2753 (2020) 415-429.

[61] C. M. Fedorov, A. Berko, Y. Matseliukh, V. Schuchmann, I. Budz, O. Garbich-Moshora, M. Mamchyn, Decision support system for formation and implementing orders based on cross programming and cloud computing, CEUR Workshop Proceedings Vol-2917 (2021) 714–748.

[62] M. Bublyk, V. Mykhailov, Y. Matseliukh, T. Pihniak, A. Selskyi, I. Grybyk, Change management in R&D-quality costs in challenges of the global economy, CEUR Workshop Proceedings Vol-2870 (2021) 1139–1151.

[63] V. Vysotska, A. Berko, M. Bublyk, L. Chyrun, A. Vysotsky, K. Doroshkevych, Methods and tools for web resources processing in e-commercial content systems, in: Int. Scientific and Technical Conference on Computer Sciences and Information Technologies, 2020, pp. 114-118.

[64] A. Berko, I. Pelekh, L. Chyrun, M. Bublyk, I. Bobyk, Y. Matseliukh, L. Chyrun, Application of ontologies and meta-models for dynamic integration of weakly structured data, in: Proceedings of International Conference on Data Stream Mining and Processing, DSMP, 2020, pp. 432-437.

[65] M. Bublyk, V. Lytvyn, V. Vysotska, L. Chyrun, Y. Matseliukh, N. Sokulska, The Decision Tree Usage for the Results Analysis of the Psychophysiological Testing, CEUR workshop proceedings Vol-2753 (2020) 458-472.

[66] A. Agresti, Analysis of Ordinal Categorical Data, John Wiley & Sons, 1984.

[67] S. Glen, Kendall's Tau (Kendall Rank Correlation Coefficient), Elementary Statistics for the rest of us, 2022. URL: https://www.statisticshowto.com/kendalls-tau/.

[68] Construction of an interval variable sequence of continuous quantitative data, 2022. URL: https://stud.com.ua/93314/statistika/pobudova_intervalnogo_variatsiynogo_ryadu_bezperernih_kilkisnih_danih.

[69] M. Bublyk, V. Feshchyn, L. Bekirova, O. Khomuliak, Sustainable Development by a Statistical Analysis of Country Rankings by the Population Happiness Level, CEUR Workshop Proceedings 3171 (2022) 817–837.

[70] Forecasting the trend of the time series by algorithmic methods, 2022. URL: http://ubooks.com.ua/books/000269/inx42.php.

[71] M. Bublyk, I. Klymus, B. Tsoniev, V. Zatkhei, Comparative Analysis of The Caloric Performance of Products for People with Cardiovascular Disease, CEUR Workshop Proceedings 3171(2022) 838–857.

[72] Statistical models of marketing decisions taking into account the uncertainty factor, 2022. URL: https://excel2.ru/articles/uroven-znachimosti-i-uroven-nadezhnosti-v-ms-excel.

[73] F. X. Diebold, Econometrics. Streamlined, Applied and e-Aware, Mc Graw Hill, Boston, 2013.

[74] N. Vlasova, M. Bublyk, Intelligent Analysis Impact of the COVID-19 Pandemic on Juvenile Drug Use and Proliferation, CEUR Workshop Proceedings 3171 (2022) 858–876.

[75] M. J. Schervish, Theory of Statistics, Springer Science & Business Media, New York, 2012.

[76] Grouping of statistical data - BukLib.net Library, 2022. URL: https://buklib.net/books/35946/

[77] O. Prokipchuk, L. Chyrun, M. Bublyk, V. Panasyuk, V. Yakimtsov, R. Kovalchuk, Intelligent system for checking the authenticity of goods based on blockchain technology, CEUR Workshop Proceedings Vol-2917 (2021) 618-665.

[78] C. Baum, An Introduction to Modern Econometrics Using Stata, Mc Graw Hill, Boston, 2020.

[79] Standard error, 2022. URL: https://ua.nesrakonk.ru/standard-error/.

[80] Standard deviation, 2022. URL: https://studopedia.su/10_11382_standartne-vidhilennya.html.

[81] K.O. Soroka, Fundamentals of Systems Theory and Systems Analysis, Kharkiv, 2004.

[82] A. Kowalska-Styczen, K. Sznajd-Weron, From consumer decision to market share - unanimity of majority? JASSS, 19(4) (2016). DOI:10.18564/jasss.3156.

[83] I.V. Stetsenko, Systems modeling, Cherkasy, 2010.

[84] S.S. Velykodnyi, Modeling of systems, Odessa, 2018.

[85] Graphic presentation of information, 2022. URL: https://studopedia.com.ua/1_132145_grafichne-podannya-informatsii.html.

[86] Y. Yusyn, T. Zabolotnia, Methods of Acceleration of Term Correlation Matrix Calculation in the Island Text Clustering Method, CEUR workshop proceedings Vol-2604 (2020) 140-150.

[87] N. Romanyshyn Algorithm for Disclosing Artistic Concepts in the Correlation of Explicitness and Implicitness of Their Textual Manifestation, CEUR Workshop Proceedings Vol-2870 (2021) 719-730.

[88] B. Rusyn, V. Ostap, O. Ostap, A correlation method for fingerprint image recognition using spectral features, in: Proceedings of the International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science, TCSET, 2002, pp. 219–220.

[89] Dataset https://www.kaggle.com/adityakadiwal/water-potability.

[90] Drinking Water Analysis Solutions https://resources.perkinelmer.com/lab-solutions/resources/docs/BRO_Drinking_Water_Analysis_Solutions_Brochure.pdf.

[91] S. Babichev, B. Durnyak, I. Pikh, V. Senkivskyy, An Evaluation of the Objective Clustering Inductive Technology Effectiveness Implemented Using Density-Based and Agglomerative Hierarchical Clustering Algorithms, Lecture Notes in Computational Intelligence and Decision Making 1020 (2020) 532-553.

[92] S. Babichev, V. Lytvynenko, V. Osypenko, Implementation of the objective clustering inductive technology based on DBSCAN clustering algorithm, in: Proceedings of Int. Scientific and Technical Conf. on Computer Sciences and Information Technologies, 2017, pp. 479-484.

[93] O. Veres, Y. Matseliukh, T. Batiuk, S. Teslia, A. Shakhno, T. Kopach, Y. Romanova, I. Pihulechko, Cluster Analysis of Exclamations and Comments on E-Commerce Products, CEUR Workshop Proceedings Vol-3171 (2022) 1403-1431.

[94] S. Babichev, M.A. Taif, V. Lytvynenko, V. Osypenko, Criterial analysis of gene expression sequences to create the objective clustering inductive technology, in: Proceedings of Int. Conf. on Electronics and Nanotechnology, 2017, pp. 244–248. doi: 10.1109/ELNANO.2017.7939756.

[95] A. Kowalska-Styczen, K. Sznajd-Weron, Access to information in word of mouth marketing within a cellular automata model. Advances in Complex Systems, 15(8) (2012). DOI:10.1142/S0219525912500804.

[96] S. A. Babichev, A. Gozhyj, A. I. Kornelyuk, V. I. Lytvynenko, Objective clustering inductive technology of gene expression profiles based on SOTA clustering algorithm, Biopolymers and Cell 33(5) (2017) 379–392. doi: 10.7124/bc.000961.

[97] I. Lurie, V. Lytvynenko, S. Olszewski, M. Voronenko, A. Kornelyuk, U. Zhunissova, O. Boskin, The Use of Inductive Methods to Identify Subtypes of Glioblastomas in Gene Clustering, CEUR Workshop Proceedings Vol-2631 (2020) 406-418.

[98] V. Lytvynenko, et. al., Two step density-based object-inductive clustering algorithm, CEUR Workshop Proceedings 2386 (2019) 117–135.

[99] I. Lurie, et. al., Inductive technology of the target clusterization of enterprise's economic indicators of Ukraine, CEUR Workshop Proceedings 2353 (2019) 848–859.