ISSN: 2688-8203
DOI: 10.33552/ACRCI.2020.02.000543
Advances in
Cancer Research & Clinical Imaging
Iris Publishers

Research Article

Copyright © All rights are reserved by Kachiashvili KJ

Indexes for Classification of Populations According to the Intensity of Cancer Diseases

Kachiashvili KJ[abc]* and Kachiashvili JK[d]

[a]*Professor, Faculty of Informatics and Control Systems, Georgian Technical University, 77, st. Kostava, Tbilisi, 380160, Georgia,*

[b]*Senior Scientific Worker, I. Vekua Institute of Applied Mathematics, Tbilisi State University, Tbilisi, Georgia*

[c]*Senior Scientific Worker, Muskhelishvili Institute of Computational Mathematics, Georgian Technical University, Tbilisi, Georgia*

[d]*Magistracy student, Faculty of Informatics and Control Systems, Georgian Technical University, Tbilisi, Georgia*

***Corresponding author:** Kachiashvili KJ, Professor, Faculty of Informatics and Control Systems, Georgian Technical University, Tbilisi, Georgia.

Received Date: March 28, 2020
Published Date: April 06, 2020

Abstract

By statistical processing of Georgian Cancer Registry data of 2015-2016, clustering (grouping) of Georgian populations was realized, according to the intensity of the cancer disease prevalence, for the purpose of priority distribution of existed resources and means in the country and for the reduction of the number of patients and improvement of the quality of treatment. Cluster analysis methods of mathematical statistics were used for the study, which was directly implemented using universal statistical software package SPSS. The concept of disease index was introduced for achieving the intruded purpose. Its several variants were determined. The study results using indexes showed that it is possible to group objectively populated areas and regions of the country by intensity of dissemination of cancer disease.

Keywords: Cancer; Disease; Disease Index; Cluster Analysis; Populated Area; Region

Introduction

It can be said that cancerous diseases are a consequence of a developed civilization. It has become especially relevant to mankind since the second half of the 20th century, when human impacts on the environment have become increasingly significant. The spread of cancerous diseases is becoming increasingly important for humanity and causing significant social, economic and environmental losses. To reduce these losses and improve the quality of life of people, many countries are increasing efforts to ensure that all cancer patients receive the best possible treatment for their ultimate well-being. Many valuable scientific researches, preventive, curative, rehabilitation and other activities are planned and implemented for this purpose. To this end, cancer registries [1-7] are being set up and developed in many countries around the world, which provide computerized databases that collect detailed information on cancer patients, including patients' treatments. These databases are then used to search for information to answer many questions related to disease and treatment. The fight against cancer is unanimously given priority worldwide. Around 14 million people worldwide get sick each year and 8 million die of cancer. The loss caused by this is alarming at present and the forecast of growing by 57% in 20 years is made. Continuous, comprehensive, and unbiased information on cancer populations is essential to monitor the spread of the disease, establish public health priorities, and evaluate the effectiveness of cancer control programs in the community. The main purpose of the population-based cancer registry is to make available individual data of all patients with cancer. The first population-based cancer registry was created in 1929 in Germany. There are several hundred such registries for today, which actively cover 21% of the world's population. An international standard for registered data has been introduced for the protection of quality, comparison to each other and for uniformity of information. Originally, cancer registries included only descriptions of illness severity, tendency, and geographic comparison. Subsequently, several registries expanded the data retrieval area to obtain patient survival rates, to determine the effectiveness of the health system. The coverage area of data registration has recently been expanded by adding clinical indicators: correct treatment, variety of care and duration of treatment. The Cancer Registry takes information from a variety of sources and, therefore, it is considered a trusted database. Registry data is used for a variety of purposes, including control and

This work is licensed under Creative Commons Attribution 4.0 License

prevention of disease outbreaks, for optimal distribution and management of financial, medicinal, human and technical resources. The purpose of the present paper is to group the populations and regions of Georgia by the intensity of the spread of cancer through statistical data processing of the Cancer Registry of Georgia, in order to prioritize the distribution of these resources and means throughout the country to reduce the overall number of patients and to improve the quality of treatment. Also, to prioritize nationwide disease prevention measures, to carry out studies for establishment of disease reasons for their further reduction, to make the research of the causal links between disease and causal factors, and so on. The proposed research methodology is of general importance as it is universal and can be used to achieve the goals set for any country or region. Many countries are working to determine the intensity of the spread of cancer and many related facts. Experts from different fields and specialties work to solve this problem. Including physicians, biologists, chemists, physicists, sociologists, specialists in mathematical statistics and computer science, and more. The results of their work are presented in numerous published reports, scientific papers, reports on international meetings, conferences, workshops and so on. Below is a brief annotation of some of this work to illustrate the problem and the actual results. Theoretical and practical results obtained for increasing the accuracy of the use of the unity of the methods "decision, discovery and classification" of artificial intelligence is considered in [8]. Of the specific examples discussed, we are interested in the problem of segmentation and classification of skin cancers. Using this as an example, authors have shown that developed by them "Topological-geometrical voting" (uses comparisons of proximity and distance) greatly improves the conventional arithmetical voting (i.e. weighted averages) method in many cases. Cancer data in India [6] were compared with life expectancy for smokers, alcohol drinkers, and overweight [9]. The association of these factors with the incidence of the disease was established by statistical methods. The paper [10] examines the median age of death of female cancer patients in the Indian city of Trivandrum. Data are taken from the Trivandrum Cancer Registry. The Kulbach-Leibler distance was used for this purpose. Different methods of selecting variables for large dimensional data are compared on the basis of lung cancer data in work [11]. A spectral-spatial classification method is proposed in work [12] for distinguishing cancer from normal tissue on a hyperspectral imaging. Tumor types are classified in paper [13], using artificial neural network based on brain imaging of astrocytoma type of different patients. A computerized decision-making system for the early detection of brain cancer is described in work [14]. In particular, the use of various statistical-based functions to calculate tissue structure is described. Based on these tissues, the segmentation of the brain tissue is classified into four categories based on the intensity of the histograms. The work [15] describes in detail the so-called "e-mail". Using the Random Forest Algorithm for Cancer Prevention as an effective, reliable and optimal classifier among many possible algorithms. The work [15] describes the use in detail of the so-called "Random Forest Algorithm" for cancer prevention as an effective, reliable and optimal classifier among many possible algorithms. The use of the Monte Carlo method shows the existence of noted properties of the considered algorithm. Thus, the summaries of the reviewed papers provide the basis for concluding that mathematical statistics data grouping methods, by their classification, allow us to solve many practical problems, including the problem of determining the prevalence of cancer, for optimal planning and implementation of measures of the preventive and administrative-organizational nature of cancer illness. A set of methods, called cluster-analysis methods, is used in mathematical statistics for separation of homogenous groups from a given set of data by a certain sign or by a set of signs. The development of cluster analysis methods began in the seventies of the last century and is still developing with great intensity. In recent years, special attention has been devoted to the development of special classification methods for big data systems. Cluster analysis methods allow us to divide the investigated objects into groups of homogeneous objects. Such groups are called clusters. Cluster analysis methods are widely used to solve practical tasks in many fields, such as industry, economics, defense, medicine, biology, agriculture, ecology and others. Cluster analysis plays an important role in data mining, pattern recognition and machine learning [16-21]. Many methods of cluster analysis and their use for solving different problems from different fields of human activity are discussed in variety of scientific works and their number is increasing day by day. As an example let's introduce some of them. The method of identification of point's clusters in multidimensional Euclidean space and its application in taxonomy is discussed in [22]. Two methods based on spatial dependencies between points are discussed: agglomerative (i.e. accumulative) and solvable (i.e. separable). The method is built on finding the nearest neighbor and then dividing it into clusters using the minimum inner sum criterion of a cluster. The procedure ensures effective reduction of the number of possible divisions. The method can be used for dichotomous dividing, but it is also well used for dividing into any number of clusters. The work [23] fundamentally addresses to the problem of cluster analysis and describes many divisive and heuristic methods. Programs developed for this purpose are also described. Monograph [24] gives a fundamental overview of the philosophy, essence, and existing methods of cluster-analysis methodologies. The application of these methods in various fields of science, including object classification, planning, engineering, and others. The appendices review books and articles on the problem under consideration, and many existing cluster-analysis software packages. The article [25,26] discusses the determination of the asymmetry of asthma using cluster analysis method. Many clinical, physiological and pathological parameters are associated with asthma. Therefore, multidimensional mathematical techniques – k means analysis, are used to identify distinct pheno-groups. In particular, k mean cluster analysis method for three different groups of asthma.

Basic Results of Investigation and their Consideration

As was mentioned above, the goal of this work is, by statistical processing of the cancer registry data, to group Georgian populated areas by intensity of cancer spread, for priority distribution of existed resources and means, with the purpose of the reduction of the total number of infected people and increasing the quality of the treatment.

Grouping of Georgian populations according to the incidence of cancer disease

The Cancer Registry data of Georgia was used to achieve this goal [7]. In particular, the study used the names of 961 settlements in Georgia with the reference to the number of population and the incidences of cancer in 2015-2016. It is clear that grouping settlements simply by the absolute number of infected people will not give the desired result to achieve the stated goal, since where there is a larger population there will always be a large number of infected people, and, in this case, small populated areas will be in unequal position in comparison with settlements with large population. To eliminate this obstacle, they use the disease intensity index to group the populations. Let us introduce the following denotations for computations of the disease intensity: a_i- the number of the population in point, and b_i- the number of patients. Then the number of infected, reduced to 100,000 inhabitants, or so called Incident Rate for i th settlement will be

$$\boldsymbol{I}_i = \frac{b_i \times 100{,}000}{a_i}, \quad i = 1,\ldots,n \tag{1}$$

Where n is the number of settlements involved in the study. The total population covered by the study is $A = \sum_{i=1}^{n} a_i$. In our case $A = 3{,}695{,}864$.

In order to divide the populated areas into three priority groups by calculated disease intensity, the direct application of the cluster analysis method [16-21], using SPSS [27], gives the results given in Table 1.

Table 1: Settlements grouping by populations incidences rates.

Disease Intensity	Settlements	Cluster Number	Cluster Center	Cluster Size
The biggest	(Tusheti)Omalo Community; (Tetritskaro) kldeisi; (Oni) Bari Community; (Oni) Shkmeri Community; (Oni) Tskhmori Community; (Kazbegi) Gudauri Community; (Kazbegi) Kobi Community; (Tsageri) SpatagoriCommunity.	3	1333.3	8
Comparatively less	Among them: c. Rustavi	1	447.71	204
Smaller than the rest	Amongthem:c.Tbilisi, c. Kutaisi, c. Batumi	2	204	714

Disease multiple index (DMI)

$$DMI = \boldsymbol{I}_i \times \frac{a_i}{A} \tag{2}$$

Disease exponential index (DEI)

$$DEI = \left(\boldsymbol{I}_i\right)^{\frac{a_i}{A}} \tag{3}$$

Disease Normalized Number (DNN)

$$DNNM = \left(\boldsymbol{I}_i\right) \times \boldsymbol{n}_i \text{ (Multiple)} \tag{4}$$

$$DNNE = \left(\boldsymbol{I}_i\right)^{\boldsymbol{n}_i} \text{ (Exponential)} \tag{5}$$

Disease Ortho-Normalized Number (DONN)

$$DONNM = \left(\boldsymbol{I}_i\right) \times onn_i \text{ (Multiple)} \tag{6}$$

$$DONNE = \left(\boldsymbol{I}_i\right) \times onn_i \text{ (Exponential)} \tag{7}$$

Where $\boldsymbol{n}_i = a_i / 100{,}000$ is the standardized number of the population, $onn_i = \left(a_i - 100{,}000\right)/100{,}000 = \boldsymbol{n}_i - 1$ is the ortho-standardized number of the population.

The values of the Cancer Indexes of the Georgian population computed by formulae (2)-(7) are given on Figure 1 and Figure 2 for the populations included in the study, respectively, collectively, i.e., the values of all indexes on a single graph and separately, that is, the values of the index per graph on a graph for visibility. As these graphs show, the indexes, and give one and the same results. From their slightly differ the results obtained by the indexes DMI, DEI and DNNM give one and the same results. From their slightly differ the results obtained by the indexes DNNM and DONNE. Index DONNM gives results different from previous indexes. So, we exclude it from further consideration (Figure 1& 2).

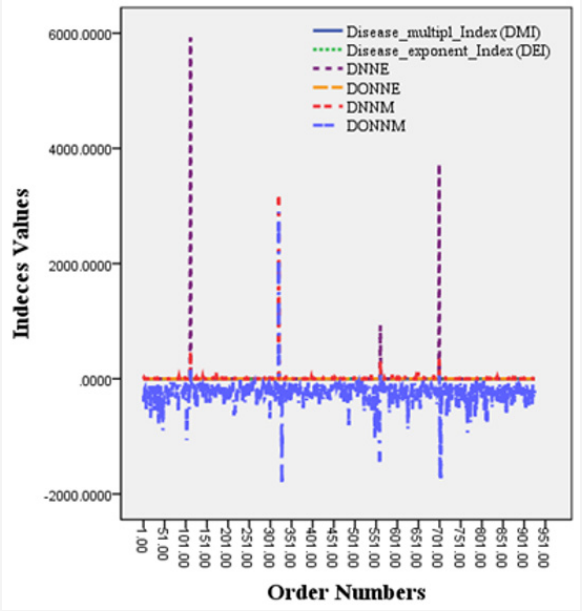

Figure 1: Values of disease indexes for Georgian populations included in the study (united).

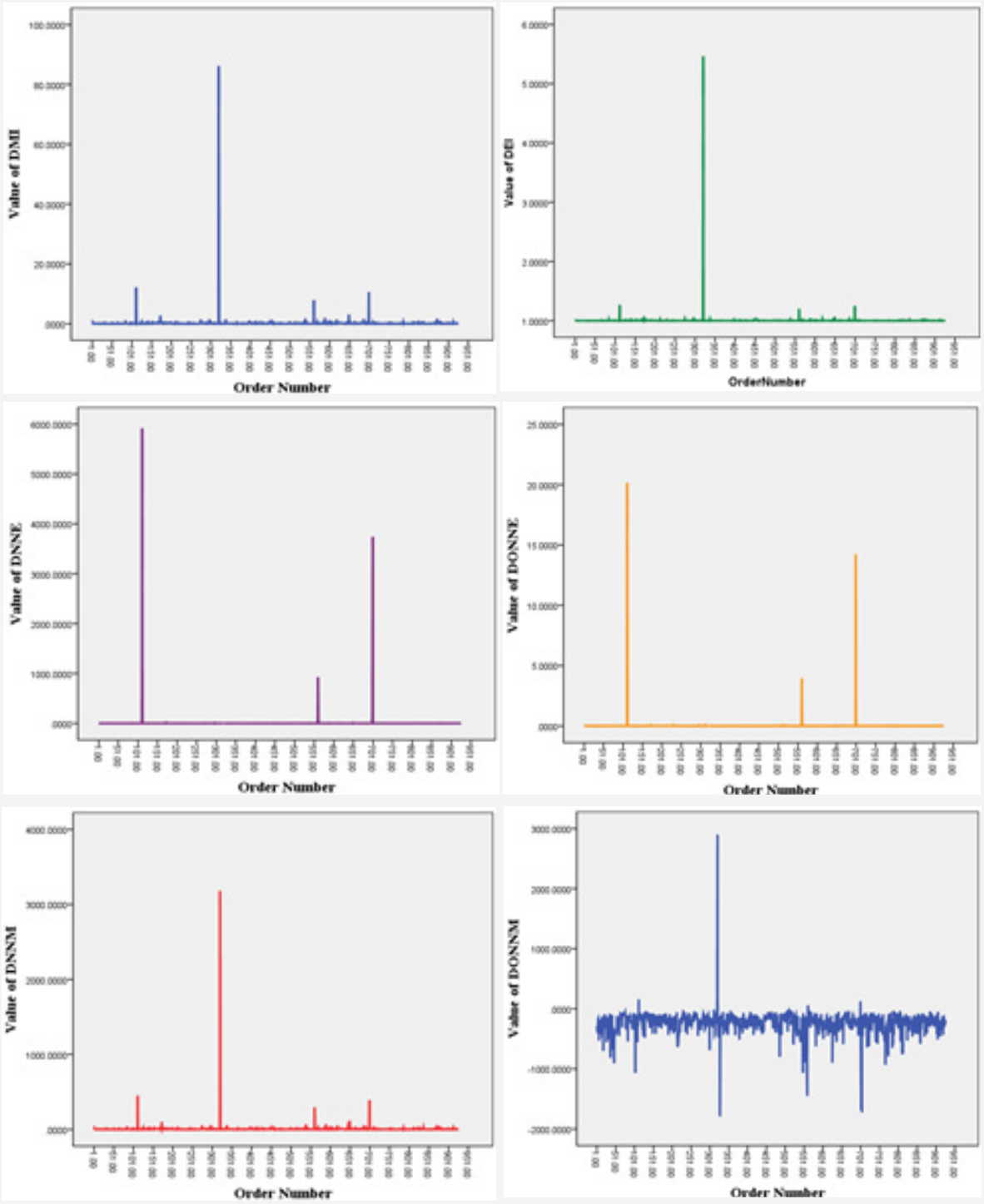

Figure 2: Values of disease indexes for Georgian populations included in the study (separated).

Grouping the regions of Georgia according to the intensity of the cancer

Taking into account economic possibility of today's Georgia, at the condition of financial capacity restrictions, for determination of the priorities of cancer disease preventive measures in order to avoid unnecessary detail, it is appropriate to not group of Georgian settlements but first of all the regions, and then - the municipalities by cancer disease intensity severity. For this purpose, clustering the data of Georgian municipalities and regions was carried out. The results of the classification of this data using multiple and exponential disease indices are given in Table 3 (Table 2&3).

Table 2: Settlemets grouping by population disease indexes.

Disease Intensity	Disease Index																			
	DMI				DEI				DNNE				DONNE				DNNM			
	SP	CN	CC	CS	SP	CN	CC	CS	SP	CN	CC	CS	SP	CN	CC	CS	SP	CN	CC	CS
Max.	T	3	86.0967	1	T	2	5.4618	1	T	3	1.78E +27	1	T	3	6.21E+24	1	T	3	3182.018	1
Med.	B, K, R	2	10.1872	3	B, K, R	1	1.2649	3	B, K	1	3	2	B, K	1	17.1757	2	B, K, R	2	376.5055	3
Min.	Oth.	1	0.1386	922	Oth.	3	1.2493	922	Oth.	2	1	923	Oth.	Oth.	0.0111	923	Oth.	1	5.1214	922

T – c. Tbilisi, B – c. Batumi, K – c. Kutaisi, R – c. Rustavi, Oth. – Other settlement points;

Max. – The biggest; Med. –Comparatively less; Min. – Smaller than the rest;

SP – Settlement Points; CN – Cluster Number; CC – Cluster Center; CS – Cluster Size

Table 3: Classification of regions of Georgia using multiple and exponential indices of disease.

Disease Intensity	Disease Index							
	DMI				DEI			
	SP	CN	CC	CS	SP	CN	CC	CS
The biggest	Tbilisi (capital)	1	0.997	1	Tbilisi (capital)	1	1.0111	1
Comparatively less	Imereti1) (Kutaisi), Adjara (Batumi)	3	0.361	2	Kahketi(Rustavi), Imereti(Kutaisi), Adjara (Batumi)	3	1.0042	3
The third largest municipality	Kahketi(Rustavi), Samrgrelo	2	0.2767	2	ZemoKartly, Samrgrelo	4	1.0027	2
The fourth largest municipality	ZemoKartly, Guria	4	0.1795	2	KvemoKartli, Guria	2	1.0013	2
The fifth largest municipality	KvemoKartli, South Georgia, Racha	5	0.0423	3	South Georgia, Racha	5	1.0005	2

We took the number of clusters equal to 5 for the purpose of relatively small, one-time investments at priority distributions of the financial and material resources. Increasing the number of clusters decreases the amount of one-time investments while decreasing the number of clusters increases the volume of one-time investments. Therefore, the selection of the number of clusters depends on the economic situation of the country and this issue is deciding by the political and economic leadership of the country. The results obtained from the separate settlements of Georgia, as well as from the municipalities and regions, show that the anti-tumor treatment and prevention measures are primarily to be implemented in city Tbilisi, then in the cities: Batumi, Kutaisi and Rustavi (in the order given), and then in the rest of the regions.

Conclusion

Statistical processing of data from the 2015-2016 Cancer Registry of Georgia was carried out by clustering (grouping) the settlements of Georgia according to the prevalence of cancer. Such grouping allows the prioritization of available resources and means to reduce the overall number of patients and improve the quality of treatment throughout the country. Mathematical statistics cluster analysis methods were used for the study. Direct use of these methods was made using universal statistical software package SPSS. The study found that grouping populations with the disease severity index that they generally use did not produce the desired results, as it suggested that anti-tumor measures should first be carried out in populated areas with only a few dozen inhabitants and the number of infected units. In order to eliminate this defect, the concept of the disease index was introduced and several options were defined. The results of the research using the indexes show that they give essentially one and the same results and can be used for objective grouping of the populations according to the intensity of the spread of the cancer, depending on the political and economic situation in the country.

Acknowledgement

None.

Conflicts of Interest

No conflict of interest.

References

1. https://www.iarc.fr/
2. https://www.encr.eu/conferences/2016-joining-forces-better-cancer-egistration-europe
3. https://www.ncbi.nlm.nih.gov/pmc/ articles/PMC5846186/

4. http://web1.sph.emory.edu/GCCS/cms/index.html
5. https://www.cdc.gov/cancer/npcr/value/registries.htm
6. http://www.ncdirindia.org
7. https://www.ncdc.ge/Handlers/GetFile.ashx?ID=1f20368c-fbf4-40f1-8fd3-6fa2ad88a11e
8. Dat TT, Hang NT, Hieu PH, Nga NT, Phuong LB et al. (2019) Optimization by voting in decision making: arithmetic versus topological voting method. Published conference program and abstract book of the International Conference on Applied Probability and Statistics (CAPS 2019), Hanoi, Vietnam, p. 40.
9. Mathew A, George PS, KM JK, Vasudevan D, James FV (2018) Transition of Cancer in Population in India. Cancer Epidemiol. 58: 111-120.
10. Jagathnath KKM, Preethi SG, Aleymma M (2018) Estimation of Cancer Barden among Women: Kullback-Laibler Divergence Approach. Souvenir & Abstract of 4th International Conference on Statistics for Twenty-First Century – 2018 (ICSTC-2018), Trivandrum, India, p. 20-21.
11. Gilhodes J, Leconte E, Boher JM, Filleron T (2019) Comparison of Variable Selection Methods for High-dimensional right Censored Data. Conference Program & Abstract of 2019 International Conference on Applied Probability and Statistics (CAPS 2019), Hanoi, Vietnam, p.37.
12. Lu Guolan, Halig LV, Wang D, Qin X, Chen ZG, et al. (2014) Spectral-spatial classification for noninvasive cancer detection using hyperspectral imaging. J of Biomedical Optics, 19(10).
13. Jain Sh (2013) Brain Cancer Classification Using GLCM Based Feature Extraction in Artificial Neural Network. International Journal of Computer Science & Engineering Technology (IJCSET), 4 (7): 966-970.
14. Sheshadri HS, Kandaswamy A (2007) Experimental investigation on breast tissue classification based on statistical feature extraction of mammograms. Computerized Medical Imaging and Graphics, 31 (1): 46-48.
15. Izmirlian G (2004) Application of theRandom Forest Classification Algorithmto a SELDI-TOF Proteomics Studyin the setting of a Cancer Prevention Trial. National Cancer Institute; Executive Plaza North, Suite 3131; 6130 Executive Blvd, MSC7354; Bethesda, MD 20852.
16. Kachiashvili KJ, Nurani B (2013) Statistical Models and Simulation by SPSS. Publisher "Alfabeta" Bandung Indonesia 353 p. (text-book)
17. Tiurin IN, Makarov AA (1998) The statistical analysis of the data on the computer. P. 528.
18. Aivazian SA, Buchstaber VM, Yenyukov IS, Meshalkin LL (1989) Applied statistics. Classification and reduction of dimentionality. Edited by prof. S. A. Aivazian, Financy i statistica Moscow p. 607.
19. Bolshev LN (1969) Cluster analysis Bull. Int. Stat. Inst. 43: 441 - 425.
20. Morrison DG (1967) Multivariate statistical methods. NY Mc Grou Hill Book Company.
21. Willmott AJ, Grimshaw PN (1969) Cluster analysis in social geography: Numerical taxonomy. Ld NY Acad Press. p. 271 - 281.
22. Caliliski T, Harabasz J (1974) A Dendrite Method for Cluster Analysis. Communications in Statistics 3(1): 1-27.
23. Kaufman L, Rousseeuw P (2005) Finding Groups in Data. An Introduction to Cluster Analysis. John Willey & Sons, p. 349.
24. Romesburg HCh (2004) Cluster Analysis for Researchers. Lulu Press North Carolina p. 333.
25. Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, et al. (2008) Cluster Analysis and Clinical Asthma Phenotypes. American Journal of Respiratory and Critical Care Medicine 17: 218-224.
26. Ye J (2017) Single-Valued Neutrosophic Clustering Algorithms Based on Similarity Measures. Journal of Classification 34:148-162.
27. Buhl A, Zofel (2001) SPSS Version 10. Einfuhrung in die Moderne Datenanalyse unter Windows. Pearson Education Deutschland GmbH p. 602.