UDC 004.415.2
DOI https://doi.org/10.32838/2663-5941/2022.4/18

Oleshchenko L.M.
National Technical University of Ukraine "Igor Sikorsky Kyiv polytechnic institute"

Movchan K.O.
Ukrainian Scientific and Research Institute of Special Equipment and Forensic Expertise of the Security Service of Ukraine

Guida O.G.
V.I. Vernadsky Taurida National University

Novak D.S.
Kyiv National University of Technologies and Design

SOFTWARE METHODS FOR ANALYSIS AND FORECASTING SUSTAINABLE DEVELOPMENT INDICATORS USING PYTHON TOOLS

Today we have the rapid development of information and communication technology, the amount of global data is exploding. An important task is to identify factors that affect the sustainable development of society through the software analysis of data from various socio-economic indicators. As for today the analysis of sustainable development is a task with a fairly high demand that is done in order to evaluate the influence of various positive and negative factors on the development of the region. With the help of such analysis one can research the economic, social, ecological and technological threats. These metrics can indicate a field having certain positive or negative dynamics. Based on the results achieved, dependencies can be found and improvements for the future can be made. The implementation of software that would automate the creation of necessary tables, performing calculations and building the charts based on the input data is a fairly necessary task.

This research is devoted to the analysis of the level of sustainable development of society. Existing methods for analysis and clustering of statistical data about the territorial living standard of inhabitants are researched. Software method of calculating the index of sustainable development of society and the components of quality of life is created. The structure of the database is modeled in the models.py file which is a Django structure file. To test the developed software, statistical data were analyzed using Python tools. The clustering module of proposed software is developed and tested. Clustering algorithms by data set size, number of clusters, by data set type are analyzed and compared.

Key words: *software, sustainable development, harmonization of society degree, clustering, statistical data, Python technologies, Django.*

Introduction. Problem statement

The full implementation of the concept of sustainable development is inextricably linked with its appropriate provision at various levels of management, the elements of which should provide for full mutual consistency to maximize obtaining a useful result of managing the components of socio-ecological and economic development. Research results show that the dynamics economic growth in recent years has been unstable, since the growth of the Ukrainian economy was determined primarily by the ability to adapt to changes external and internal institutional and economic conditions, rather than the formation of sustainable internal principles of balanced development. At present, the entire world community is facing an aggravation of not only economic, but also social problems against the backdrop of deteriorating environmental conditions.

As for today the analysis of sustainable development is a task with a fairly high demand that is done in order to evaluate the influence of various positive and negative factors on the development of the region.

With the help of such analysis one can research the economic, social, ecological and technological threats. These metrics can indicate a field having certain positive or negative dynamics. Based on the results achieved, dependencies can be found and improvements for the future can be made. The implementation of software that would automate the

creation of necessary tables, performing calculations and building the charts based on the input data is a fairly necessary task.

Related research

The research is based on the sustainable development concept that is derived from Volodymyr Vernadsky's works on the noosphere [1]. Both theory and practice have proven that at the turn of the century the teachings about noosphere turned out to be a necessary basis for developing the concept of continuous ecological, social and economic development [2–4].

Methodology of evaluating and analyzing sustainable development covers the model of sustainable development, which is a cross-field generalization of the models known from natural, economic and social science fields, and the ways of utilising formal statistical methods and methods of expert evaluation for analyzing the processes of sustainable development.

The main goal of the article is to developed the software method for calculation of the coefficient of sustainable development, analysis of factors of the degree of harmonization of society and speed up time of processing statistical data about the territorial living standard of inhabitants and to obtain more accurate results compared to existing methods.

Mathematical basis of the proposed method

We will characterize the process of sustainable development with two main components: safety (C_{sl}) and quality of life (C_{ql}), and the generalized measurement of sustainable development will be defined by the following quaternion:

$$\{Q\} = jw_{sl}C_{sl} + w_{ql}\overline{C_{ql}}\left(I_{ec}, I_e, I_s\right). \qquad (1)$$

Quaternion $\{Q\}$ contains an imaginary balanced scalar faction $jw_{sl}C_{sl}$ that describes the safety of life and a balanced real vector faction that describes the quality of life in a space with three dimensions: economic (I_e), ecological (I_{ec}) and the dimensional of social institutes (I_s). In addition, j gains the value of a real unit for a regular state of the development of society while $C_{sl} > 0$ and the value of an imaginary unit when the society enters a state of conflict ($C_{sl} = 0$).

Balance coefficients w_{sl} and w_{ql} in the formula (1) are used in order to equalize the scales of safety and quality of life components (in case of evaluating sustainable development of Ukraine's regions $w_{sl} = w_{ql} = 1$).

Index of Sustainable Development is a quantity measurement of sustainable development, that includes safety and quality of life and for the case of $C_{sl} > 0$ is calculated as the norm of quaternion $\{Q\}$:

$$\{Q\} = \sqrt{w_{sl}^2 C_{sl}^2 + w_{ql}^2\left(I_{ec}^2 + I_e^2 + I_s^2\right)}. \qquad (2)$$

For every region the euclidean norm of quality of life radius-vector $\overline{C_{ql}}$ will be given in the following form:

$$\overline{C_{ql}} = \sqrt{I_{ec}^2 + I_e^2 + I_s^2}. \qquad (3)$$

Than the quantity measurement of quality of life will be defined as the length of a projection of this vector onto an vector with the coordinates of (1; 1; 1):

$$C_{ql} = \sqrt{I_{ec}^2 + I_e^2 + I_s^2} \cdot cos\left(\alpha\right). \qquad (4)$$

Deviation angle α of radius-vector $\overline{C_{ql}}$ from the vector (1, 1, 1) is defined by the values of the measurements I_{ec}, I_e, I_s in the following way:

$$\alpha = arccos\frac{I_{ec} + I_e + I_s}{\sqrt{3} \cdot \sqrt{I_{ec}^2 + I_e^2 + I_s^2}}, \qquad (5)$$

$$0 \le \alpha \le arccos\frac{1}{\sqrt{3}}.$$

Accordingly, the length of a projection of the radius-vector $\overline{C_{ql}}$ onto an ideal vector (1, 1, 1) describes the quality of life, while the orientation of the vector $\overline{C_{ql}}$ inside the coordinate space (I_{ec}, I_e, I_s) describes the measure of the "harmony" of sustainable development.

Vector $\overline{C_{ql}}$'s equal distancing from every one of coordinates will correspond to the most harmonic development while closing in to one of the coordinates will indicate a priority in development in the corresponding dimension and negligence toward the other two. The value $G = 1 - \alpha$ will be named the level of harmony of sustainable development. It will grow when G is closing to 1 and drop with G closing to 0. Accordingly the Quality of life component is an integrated value that simultaneously includes all three dimensions of sustainable development and the interconnection between three inseparable fields of society's development: economic, ecological and social. Level of harmony of sustainable development is the balance between its economic, ecological and social dimensions.

The values of sustainable development dimensions that were used to define the Quality of life component should be based on the data from a wide spectrum of phenomena of various nature. Additionally these values must be integral, meaning they must describe a certain aspect of human life as a coherent system. In order to evaluate the quantity measurements of the sustainable development dimensions we will use the

principles of building a hierarchical system of values and indexes defined as L_1-norms:

$$I_i = \sum_{j=1}^{n} w_j x_{i,j},\ i = \overline{1,m},\ \sum_{j=1}^{n} w_j = 1 \qquad (6)$$

in the space of values $X^1 \times X^2 \times \ldots \times X^m$, that describe economic, ecological and social development of every i-th region. Balance coefficients w_i in the formula (6) are defined by expert evaluation.

Using formula (6) requires the harmonization of various data, including both measurement units and value ranges [5]. That's why if greater values of a measurement X^i correspond to a better state of sustainable development, then we use a logistical norm for measurement values according to formula:

$$C_{norm}(x_{i,j}) = \left(1 + e^{\frac{a - x_{i,j}}{b}}\right)^{-1}, \qquad (7)$$

where the parameters a and b are calculated as an average value and standard divergence of the sample of regions being analyzed.

In the opposite case if greater values of X^i correspond to worse state of the sustainable development, then a value inverse to one calculated through formula (7) is used:

$$C_{norm}(x_{i,j}) = 1 - \left(1 + e^{\frac{a - x_{i,j}}{b}}\right)^{-1}. \qquad (8)$$

The total influence of threats on different regions of Ukraine will be evaluated using a Safety of life component C_{sl} as a part of the Index of sustainable development used in the formula (1).

To every region j a corresponding vector:

$$Tr_j = (t_i^j),\quad j = \overline{1,n} \qquad (9)$$

will be assigned with its coordinates $t_i^j \in [0,1], i = \overline{1,n}$ describing the level at which every threat manifests.

The value of measurements describing the occurrence of said threats are harmonized using the formulas (7)–(8) in a way where the greater threat levels have their correspondent values close to 1 [1–5].

To organize the decision-making process aimed at sustainable economic, socio-natural development of the region, it is necessary to develop a methodology for calculating the integral indicator of sustainable development of the regional economy, where the main indicators would be closely linked to the targets and priority areas for the development of the strategic plan of the region.

When calculating indicators of sustainable development, there are two methodological approaches that differ in structure and applied principles of construction. In the first approach, all calculated indicators of the system reflect certain aspects of sustainable development, that is, they distinguish the following subsystems of indicators: economic, social, environmental and institutional. In this case, they usually consider various options for issues, problems, tasks that are submitted for consideration, which may not even have a quantitative characteristics, but only a descriptive answer. In the second approach, an integral indicator is constructed that indicates the degree of sustainability of economic and social and environmental development. With the growth of such an aggregated indicator, the economy of the region takes the path that brings it closer to sustainable development, and with its decrease or with a negative value, the movement occurs in the opposite direction to the destructive type of development.

Statistical data for research

For completing analyzing the sustainable development of Ukraine's regions the task was complemented with a .osd file that contains the data on all Ukraine's regions in two years (2017 and 2018). This file contains such indicators as: Level of housing needs satisfaction, The number of infected with HIV, Integrated atmospheric pollution index averaged on the population of observed towns etc. All this data had to be read from the file and develop a procedure that would calculate the model of sustainable development (composite indicator that is calculated based on measurements) and harmony level using the programming tools of Python.

Using programming tools of Python we represent the sustainable development index and harmony level (visualize the data) by the regions and find the correlation chart between all the indicators (by the use of corresponding interface buttons). Additionally a sample of statistically influential indicators ($r > 0.7$ and $r < -0.7$) had to be made and a conclusion about their corresponding levels of connection had to be formulated. Next we had to perform a regression analysis of statistically important indicators (independent variables $X_1, X_2 \ldots, X_p$) influence on the sustainable development index (dependent variable Y) and separately on the harmony level (also as a dependent variable). Develop according regression mathematical models for all 27 regions of Ukraine. Also a database for storing input data and generated results of input files analysis had to be created. We need to create a database to store the input files and the generated analysis results of the input files. Import the results of calculating the sustainable development

index (I_{sd}) and harmony level (G) in .csv and xls formats for all 27 regions of Ukraine.

The proposed software method and software modules description

Python is one of the most widely used programming languages thanks to it being easy to learn, well-designed and flexible, making it practically a perfect programming language. Django is an immensely popular and fully functional server-side web framework written in Python. This framework can work with any client-side platform and can deliver content in almost any format (including HTML, RSS-channels, JSON, XML). Django includes dozens of additional functions that fulfil user authentication, site mapping, content administration, RSS etc. Django uses a component-based architecture (every component of the architecture is independent from the others and thus can be replaced or altered in case of such necessity). Django framework is extremely well suited for high traffic workloads. Django uses the Do not Repeat Yourself (DRY) principle, thus avoiding unnecessary code repeats, reducing the overall quantity of code [6–7].

For the data output in comfortable format, the output of tables and charts the following instruments had to be used. Highcharts is a library for chart creation that is written in Javascript. It allows adding interactive and animated charts to the website or the web application easily. At the moment charts include a large variety of line graphs, pie charts, column charts and many other types of graphs. Also a standard Python set of tools for mathematical operations and data analysis was utilised. The proposed software contains the following modules:

1. Module for reading and processing the information from the file, which contains the data on 27 territorial units of Ukraine.
2. Module performing the procedure of calculating the sustainable development index model (composite indicator that is calculated based on measurements) and harmony level.
3. Software modules for the graphical representation of sustainable development index and harmony level (data visualization) by the regions and finding the correlation chart between all indicators (by the use of corresponding interface buttons).
4. A module performing regression analysis of statistically important indicators influence on the sustainable development index and separately on the harmony level and building according regression mathematical models for all 27 regions of Ukraine.
5. A module for storing input data and generated results of input files analysis.
6. A module for importing the results of calculating the sustainable development index (I_{sd}) and harmony level (G) in .csv and .xls formats for 27 regions of Ukraine into the database.

For the better understanding of the software structure a visual representation of all the modules and components of the project is provided. Here is a dependency scheme of the proposed software front-end (chart visualization part) (fig. 1).

One of the main parts in implementing software is the database. The structure of the database is modeled

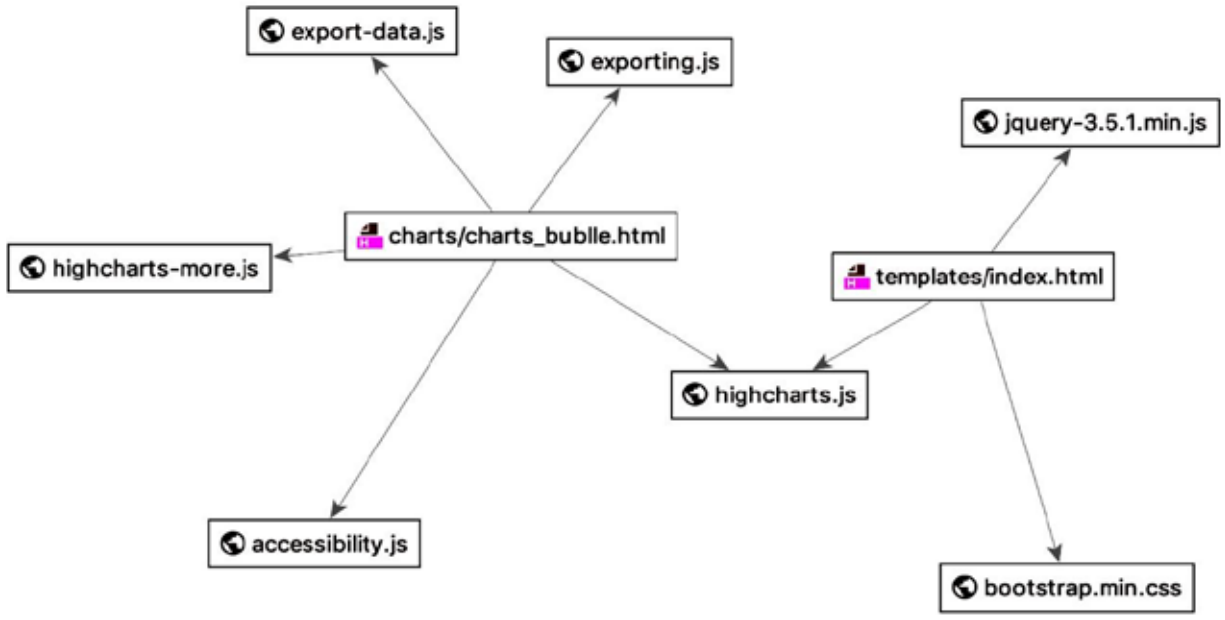

Fig. 1. Visualization module

in the models file (models.py) which is a Django structure file. After modelling the database tables that are correspondent to the models are created automatically via migration.

Data reading and processing module. In order to use the application one has to download the .osd data file containing the information on 27 territorial units of Ukraine. This functional possibility is taken response of by the method export_data in the file views.py. Here we receive the data and perform the calculations of all the key indicators. After the data has been processed the calculation of the indicators has to be performed. After the indicators have been calculated we have to provide the possibility of saving this data into the database.

Thus we have the functionality for the page "Upload a file", which can be used to upload and save the data and also show the resulting values (fig. 2).

Data visualization module. For outputting the vulnerability correlation chart for the harmony level and sustained development index we have to consult the file charts.py and the method chart_garmonizacii where the process of calculating the correlation coefficient occurs. This coefficient is later loaded into the corresponding chart (fig. 3) where the user can review the received data.

Data analysis module. For implementing the regressive model we determine the correlation coefficient for every region and every type of threats [8]. For this task we use the method charts_buble from charts.py. For a start we use a loop to look through all the regions and get the necessary indicators from every one of them, and then use the data from these coefficients to calculate the correlation coefficient.

According to the analysis of the Sustainable Development Index for Ukraine, we will perform clustering of regions of Ukraine on this indicator into five clusters (Highest level of sustainable development, High level of sustainable development, Average level of sustainable development, Level of sustainable development below average, Low level of sustainable development) (fig. 4).

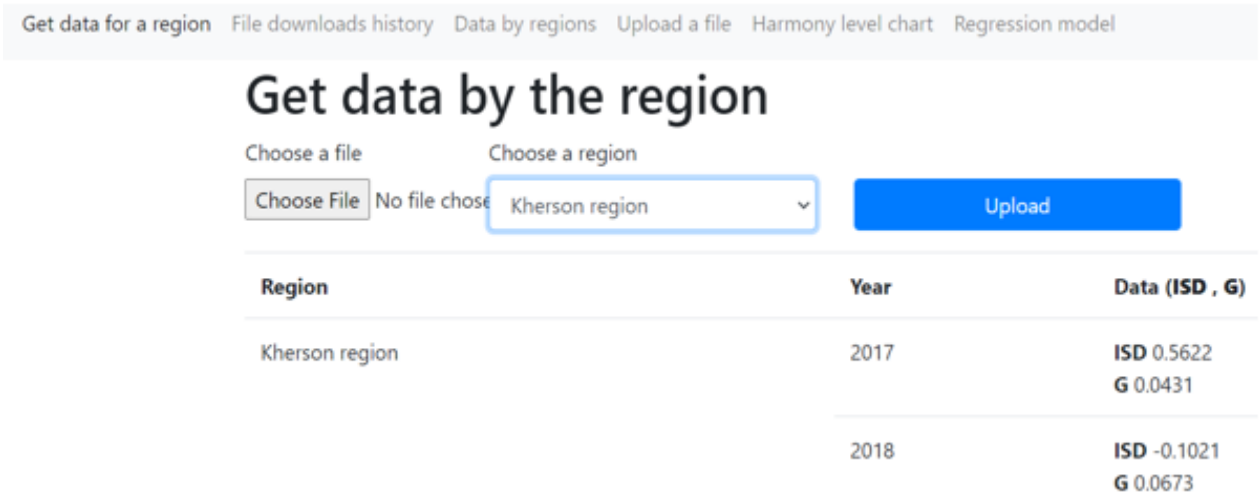

Fig. 2. View of the resulting data

	Vulnerability to threats						
	Social type		Economic type	Ecological type			
	Corruption	Healthcare index	Economic well-being	Climate change	Air pollution	Load on water resources	Technological threat
Sustained development index	-0.0165	-0.2192	-0.2236	0.1119	-0.0187	-0.2011	0.2444
Harmony level	-0.0153	-0.2029	-0.207	0.1036	-0.0174	-0.1861	0.3784

Fig. 3. Vulnerability to threats

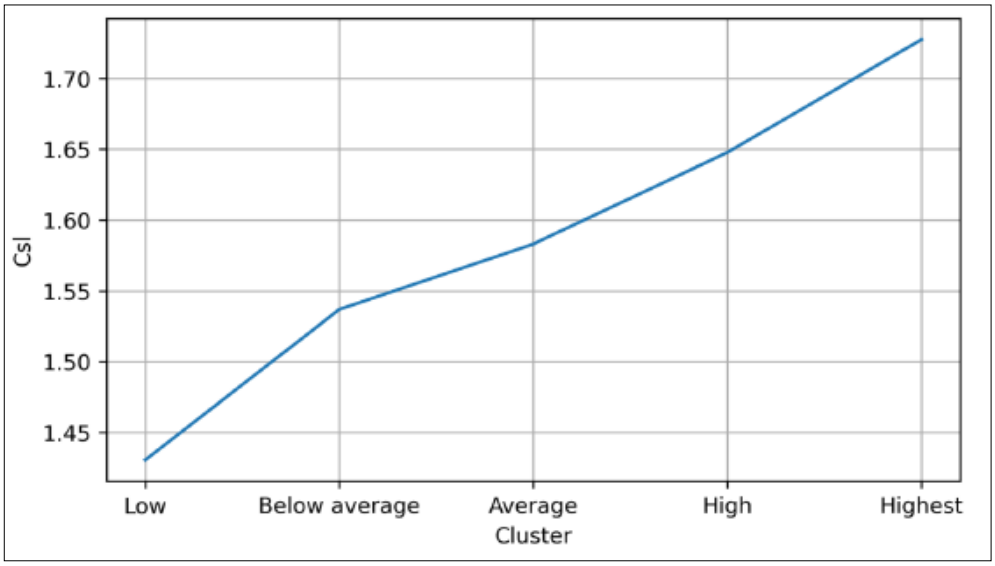

Fig. 4. Mean values of the life safety components for clusters, 2018

```
kmeans = KMeans(n_clusters=5, init='k-means++')
clusters_csl_2018 = pd.Series(kmeans.fit_predict(data_2018.loc[:, ['Csl']]), index=data_2018.index, name='cluster_num')
pd.concat([data_2018['region_name'], clusters_2018, clusters_csl_2018.rename('cluster_csl')], axis=1)
```

We analyze the main parameters that affect the sustainable development of regions to build a scattering diagram of the dependence of the Sustainable Development Index on the degree of harmonization. We see that in both years in clusters with a higher index of sustainable development, the average index of life safety is also higher. But if we compare clusters by sustainability index and security component, they are different. Kyiv city, for example, no longer has its own cluster – the clusters are more uniform. Both trends are repeated in 2017 and 2018.

In Kyiv city, the economic component significantly outperforms other components of quality of life. If we talk about the main components of the index of sustainable development, the quality of life index has a greater impact on it, but the positive correlation between the two main components is significant.

For example, in 2018, both components have less impact (fig. 5). In the correlation tables of the Sustainable Development Index and the Vulnerability Index we can see that in 2017 the largest inverse correlation has the provision of the population with doctors, and in 2018 – the average life expectancy.

To get the regression model we navigate to the /charts-bubble/ page where we review the correlation chart of the regions' vulnerability to threats and the regression models for every threat category are presented (fig. 6). To use the required chart for every threat type, we have to use the side buttons.

Research results

Functioning of clustering algorithms depends on the mathematical & logical procedure which they use to solve a problem, and also on the input dataset [9–10]. Hierarchical clustering algorithms are not suitable for large dataset due to time complexity. *K*-means can form clusters very effectively and faster than most algorithms. For our research the four clustering algorithms (*k*-means, Hierarchical Clustering (HC), Self-Organization Map (SOM) algorithm, Expectation Maximization (EM) clustering algorithm) are compared according to the following factors: the size of datasets, number of clusters, dataset type. The performance of different algorithms for different *k* is compared in order to test the performances that are related to *k*. To compare Hierarchical Clustering with other algorithms, the hierarchical tree is cut at two different levels to obtain corresponding numbers of clusters. *K*-means and EM algorithms have less quality (accuracy) than others. All the algorithms have some ambiguity in some noisy data to be clustered. The small dataset is extracted as a subset of the huge data set. The quality of EM and *k*-means algorithms becomes very good when using a huge data set. The other two algorithms Hierarchical Clustering and SOM show good results when using a small data set.

Partitioning algorithms (EM and *k*-means) are used for huge dataset while hierarchical clustering algorithms are used for small data set. For *k* from 3

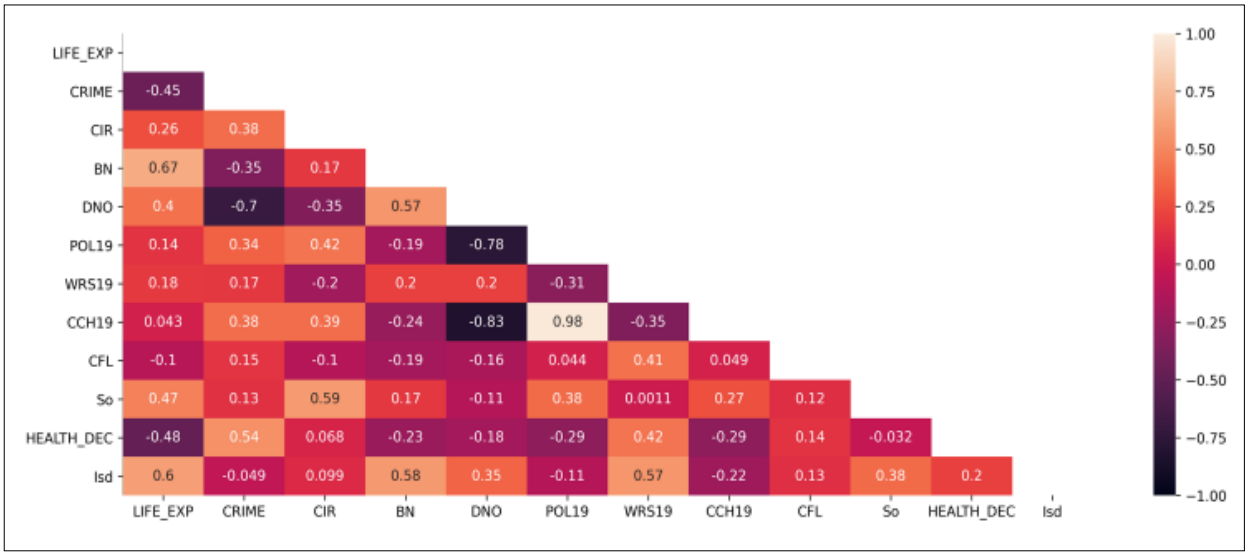

Fig. 5. Correlation coefficients between sustainable development indicators and vulnerability to the impact of threats, 2018

Fig. 6. Regression model for corruption indicator

to 8 on average we get 3–8% better result. After analyzing the results of testing the clustering algorithms and running them under different factors and situations, the followiing conclusions obtained: as the number of clusters *k* becomes greater, the performance of SOM algorithm becomes lower. The performance of *k*-means and EM algorithms is better than hierarchical clustering algorithms. The quality of *k*-means and EM algorithms become a very good when using huge data set.

In our research we used the *k*-means++ algorithm to choose the initial values, or the initial cluster centroids, for *k*-means, because in some cases, if the initialization of clusters is not appropriate, *k*-means can result in arbitrarily bad clusters. *K*-means++ specifies a procedure to initialize the cluster centers before moving forward with the standard *k*-means clustering algorithm.

Using the *k*-means++ algorithm, we optimize the step where we randomly pick the cluster centroid. We are more likely to find a solution that is competitive to the optimal *k*-means solution while using the *k*-means++ initialization.

Conclusions and future work

Existing methods for analysis and clustering of statistical data about the territorial living standard of inhabitants are researched. Software method of calculating the index of sustainable development of society and the components of quality of life is created. To test the developed software, statistical data were analyzed using Python tools. The clustering module of proposed software is developed and tested.

The four clustering algorithms (*k*-means, Hierarchical Clustering, Self-Organization Map algorithm, Expectation Maximization clustering algorithm) are compared according to the following factors: the size of datasets, number of clusters, dataset type. For number of clusters *k* from 3 to 8 on average we get 3–8 % better result using *k*-means algorithm.

Through proposed software method, problems can be clearly identified and solved as soon as possible, so as to provide a better living environment and living standard for the people. Developed software for analyzing the sustained ment of Ukraine's regions allows to analyze the data and receive conclusive results.

References:

1. Вернадский В. И. Несколько слов о ноосфере. *Успехи современной биологии*. 1944. No 18, вып. 2. С. 113–120.
2. Zgurovsky M. Z. Sustainable development global simulation: Opportunities and threats to the planet. *Russian Journal of Eearth Sciences.* 2007. Vol. 9, ES2003, doi: 10.2205/2007ES000273.
3. Аналіз сталого розвитку – глобальний і регіональний контексти: у 2 ч. Ч. 2. Україна в індикаторах сталого розвитку. Аналіз – 2009. Виконавці: А.О. Болдак, С.В. Войтко, І.М. Джигирей та інші : наук. кер. М.З. Згуровський. К. : НТУУ «КПІ», 2009. 200 с.
4. Аналіз сталого розвитку – глобальний і регіональний контексти: у 2 ч. Ч. 1. Глобальний аналіз якості та безпеки життя людей. Аналіз – 2009. Виконавці: А.О. Болдак, С.В. Войтко, І.М. Джигирей та інші: наук. кер. М. З. Згуровський. К. : НТУУ «КПІ», 2009. 280 с.
5. Згуровский М.З., Болдак А.А., Ефремов К.В. Интеллектуальный анализ и системное согласование научных данных в междисциплинарных исследованях. *Кибернетика и системный анализ.* 2013. No 4. С. 62–75.
6. Django documentation. URL: https://docs.djangoproject.com/en/2.2/
7. Advantages and Disadvantages of Python Programming Language. URL: https://medium.com/@mindfiresolutions.usa/advantages-and-disadvantages-of-python-programming-language-fd0b394f2121
8. Linear Regression in Python. URL: https://realpython.com/linear-regression-in-python/
9. Sushant Bhargav, Mahesh Pawar. A Review of Clustering Methods forming Non-Convex clusters with, Missing and Noisy Data. *International Journal of Computer Sciences and Engineering.* Mar 2016, Vol. 4 (3), P. 39–44.
10. Jain A., Dubes R. Algorithms for Clustering Data. Prentice Hall, IGI Global, 2012, pp. 43-62.

Олещенко Л.М., Мовчан К.О., Гуйда О.Г., Новак Д.С. ПРОГРАМНІ МЕТОДИ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ ПОКАЗНИКІВ СТАЛОГО РОЗВИТКУ З ВИКОРИСТАННЯМ ІНСТРУМЕНТІВ PYTHON

Сьогодні спостерігається стрімкий розвиток інформаційно-комунікаційних технологій та збільшення обсягів великих даних. Важливим завданням є виявлення факторів, що впливають на сталий розвиток суспільства шляхом програмного аналізу даних різних соціально-економічних показників. На сьогоднішній день аналіз сталого розвитку є досить затребуваним завданням, яке виконується з метою оцінки впливу різних позитивних і негативних факторів на розвиток регіону. За допомогою такого аналізу можна досліджувати економічні, соціальні, екологічні та технологічні загрози. Ці показники можуть вказувати на поле з певною позитивною чи негативною динамікою. На основі досягнутих результатів можна знайти залежності та пропонувати покращення показників на майбутнє. Актуальним завданням є впровадження програмного забезпечення, яке дозволило б автоматизувати створення необхідних таблиць, виконання розрахунків та побудову діаграм на основі вхідних даних.

Дане дослідження присвячене аналізу рівня сталого розвитку суспільства. Досліджено існуючі методи аналізу та кластеризації статистичних даних. Створено програмний метод розрахунку індексу сталого розвитку суспільства та складових якості життя. Структура бази даних моделюється у файлі models.py, який є файлом структури Django. Для тестування розробленого програмного забезпечення були проаналізовані статистичні дані за допомогою інструментів Python. Розроблено та апробовано модуль кластеризації запропонованого програмного забезпечення. Проаналізовано та порівняно алгоритми кластеризації за розміром наборів даних, кількістю кластерів, за типом набору даних.

Ключові слова: *програмне забезпечення, сталий розвиток, ступінь гармонізації суспільства, кластеризація, статистичні дані, технології Python, Django.*