Вісник Тернопільського національного технічного університету
https://doi.org/10.33108/visnyk_tntu
Scientific Journal of the Ternopil National Technical University
2022, № 1 (105) https://doi.org/10.33108/visnyk_tntu2022.01
*ISSN 2522-4433. Web: **visnyk.tntu.edu.ua***

UDC: 004.67:519.25

DISTANCE MEASURES-BASED INFORMATION TECHNOLOGY FOR IDENTIFYING SIMILAR DATA SERIES

Anastasiia Baturinets

Oles Honchar Dnipro National University, Dnipro, Ukraine

Summary. *The aim of the work is to develop and implement a technology for identifying similar series, and to test on series of data represented by hydrological samples. The subject of the study is the methods and approaches for identifying similar series. The object of the study is the process of identifying similar series, which are represented by certain indicators. The task is to propose and implement distance measures, where one of them takes into consideration the similarity between the values of the series and their relationship, and another is based on a weighted Euclidean distance taking into account the need to actualize the values that are the most important under certain conditions of the task; to implement a technology to find similar series represented by certain indicators values; to obtain a more resilient solution, to implement a procedure for determining a set of similar series based on the results obtained for each individual distance; the results should be analyzed and the conclusions have to be drawn dealing with practical application of the technology. The following methods were used: statistical analysis methods, methods for calculating distances, and similarity between data series. The following results were obtained: the technology for similar data series detection has been implemented; two distance measures were proposed and described as a part of the technology implemented; a procedure for determining a set of similar rows was implemented that was based on the obtained distances calculation. The scientific novelty of the research under discussion involves: Euclidean weighted distance was described and applied taking into account the actuality of data series values; a new measure of distance has been described and applied that allows both the degree of similarity between the values of the series and their correlation to be taken into account, as well as a technique has been developed for determining similar series from a set of selected distance measures. The practical importance of the developed and implemented technology consists in the following possibilities application to data series of different applied fields: conducting an assessment and identifying some similar series, in particular as an intermediate step in the analysis; in addition, the proposed distance measures improve the quality of identifying similar data series. In our further research, we plan to investigate the possibilities of lengthening the data series and filling in the gaps with values from other series defined as similar ones.*

Key words: *distance measures, similarity of numerical series, LCS, DTW, TSD, similarity of data series, hydrology.*

https://doi.org/10.33108/visnyk_tntu2022.01.128 *Received 18.01.2022*

Problem statement. A great number of problems from different applied fields deal with identifying the objects' similarities including those whose solution is an intermediate stage of the analysis. All the problems of image classification, clustering, identification, etc. are based on the task of the objects' similarity determination.

The problems of classification and clustering are to divide a set of objects into some groups or into classes. Though, taking into account some special features of different applied problems, the use of the above-mentioned approaches has certain preconditions and disadvantages. For instance, classification as an approach to study with a teacher, requires having some specified data for study and analysis of the obtained results quality. Clustering, in contrast to classification, does not require any a priori knowledge of an object belonging to a certain class, but it is possible to obtain different results by using various algorithms and distance measures, though their interpreting and the obtained groups quality analysis depend mostly on the researcher.

Nevertheless, though the use of the sets of algorithms to determine the series similarity has provided the obtained solutions stability, it has increased the complexity of the calculations considerably.

 Corresponding author: Anastasiia Baturinets; e-mail: baturinets.anastasiya@gmail.co

Analysis of the known results of the research. Some basic approaches in the problems of time series clustering have been taken into consideration in the paper [1]. Some results of the complex study of clustering, clustering algorithms, distance measures, methods of the clustering results assessment, etc. in particular, have been presented by the authors [2]. Some impact of distance measure on the clustering result has been studied in the paper [3].

Thus, distance measures and similarities are important components of object grouping algorithms implementation. Euclidean distance and its variations, namely Manhattan distance etc., are the simplest and the most widely spread in calculations. For instance, the weighted Euclidean distance taking into account the importance of each emergency characteristic under railway transport accidents clustering conditions was described in the paper [4].

The dynamic time warping algorithm should be distinguished among the more complex ones, which nowadays is the most popular distance measure, as a great number of papers have been devoted to the use and analysis of both its classic version, and the numerous modifications of the algorithm, and some of them can be seen in the papers [5–10].

Along with the dynamic time warping algorithm, the similarity measure like the longest common subsequence (LCS) is used quite frequently, which, unlike most distance and similarity measures, is more resistant to noise. LCS algorithm description, its modifications, and application examples can be found in the papers [11–15].

The study has been carried out by the authors in the paper [16], where the efficiency of use of different measures of similarity and distance with different approaches to time series presentation has been analyzed. The review and analysis of metrics to assess the intelligent information systems have been made in the paper [17], namely, the metrics used for the assessment of the systems of various nature and their elements are shown in table 1.

Thus, the problems of the paper under discussion have not proved any reasonable use neither classification algorithms nor clustering ones as, firstly, there are no defined data to be studied, secondly, the series division into classes cannot solve the set problem, and will require some further analysis.

Paper purpose. The following problem must be solved: determine the most similar series of indices, not more than the user has specified for certain series represented by certain characteristics. Therefore, the main purpose of the paper under discussion is to study the methods and the ways of similar data series determination.

To achieve the goal set one should develop, describe and implement the technology of the certain indices similar series defining by solving the following problems:

1) choose and implement the distance measures which are calculated between the data series values. To solve this problem it is necessary: to implement the known distance measures that are used for numerical data series; propose and implement the distance measure which takes into account both the value of the correlation coefficient and the distances between the series; describe and implement the weighted Euclidean distance making more influential the values which are more important under certain conditions of the problem;

2) find a set of similar series, and to do this it is necessary to implement the procedure of a similar series set determination on the basis of the obtained results by each separate distance;

3) provide the presentation of the technology results as curves, maps, tables, values of calculated distances, and assessment, which makes it possible to make the analysis of the obtained results.

Basic material. First of all, we should determine which distance measures will take part in the calculations.

Dynamic time warping (DTW) is a widely-used method to find some similarities between two series. The characteristic feature of the method is, that the series are defined as similar ones when they are similar by shape but have some warping along the time axis.

There are a great number of modifications of the DTW algorithm, though the general sequence of calculations can be presented as follows: a distance matrix is being constructed between the two series (Euclidean distance is used in the paper under discussion) → a deformation matrix is being constructed using the developed distance matrix → a deformation path is being constructed according to the deformation matrix → being based on the found path, the distance is calculated using the formula:

$$DTW(Q,S) = \min\left(\frac{\sum_{i=1}^{N} d(w_i)}{K}\right), \quad (1)$$

where Q – is the series, for which similar ones are being looked for; S – data sequences which are compared; $d(w_i)$ – the value of the certain element of the path on i – step; K – is the number of elements of the constructed path.

Besides, a well-known method of the longest common subsequence (LCS) is used in the study, which takes into account the advantages of dynamic programming and makes it possible to determine some similar series.

The characteristic feature of the method LCS is that it is more resistant to noise than the majority of the distances used. The classic version of LCS constructs a matrix taking into account the similarity of two data series. Let's take two series Q and S of length n and m respectively, then the recurrent function of similarity will look like the formula [18]:

$$LCS(i,j) = \begin{cases} 0, & if \;\; i = 0 \; or \; j = 0 \\ 1 + LCS[i-1, j-1] & if \; |Q_i - S_j| < \varepsilon \\ \max(LCS[i-1, j], LCS[i, j-1]) - & otherwise \end{cases} \quad (2)$$

where $i = \overline{1,T}$; $j = \overline{1,M}$; T, M – length of the correspondent series; ε is found by the formula:

$$\varepsilon = (\max(Q) - |\min(Q)|) \times p, \quad (3)$$

where p – the parameter from the range [0;1].

The higher value of the parameter p, the larger range of deviation is, and the default value $p = 0.025$ was determined.

The distance value was calculated by the formula:

$$LCS_{dist}(Q,S) = \min(n,m) - LCS(Q,S) \quad (4)$$

Euclidean distance was found by the formula:

$$d = \sqrt{\sum_{i=1}^{N} (Q_i - S_i)^2}. \quad (5)$$

Manhattan distance was calculated by the formula:

$$d = \sum_{i=1}^{N} |Q_i - S_i|. \quad (6)$$

The weighted Euclidean distance taking into account the valid values was calculated by the formula:

$$d = \sqrt{\sum_{i=1}^{N} w_i \times (Q_i - S_i)^2}, \quad (7)$$

where w_i was calculated by the formulae:

$$w_i = \frac{1}{N - t + 1} \quad (8)$$

or

$$w_i = \frac{t}{N}, \quad (9)$$

where N – is the number of values of the series; $t = \overline{1, N}$.

Thus, the coefficient w_i will determine the force of impact of each constituent of the distance, and will pay more attention to those values which are more important for a certain applied problem solution.

In the problem of data series lengthening using the similar series indices to fill in the absent values in a certain series under study, the calculations of the distance values as well as the calculation w_i are performed from the left to the right along the time axis, and it does not matter, the data of what period are to be restored.

The formula for calculation w_i will be chosen in the following way:

– if the period of the series lengthening is prior to the period of joint observations (fig. 1), then w_i will be calculated by the formula (8), and so the value w_i for each pair of values of the compared series will be getting lower in the direction from the point «a» towards the point «b». The highest value w_i will be obtained by the pair of values of the series which correspond to the point «b», and the lowest one – to the point «a».

– if the period of the series lengthening is after the period of joint observations (fig. 1), then w_i will be calculated by the formula (9), and so the value w_i for each pair of values of the compared series will be getting lower in the direction from the point «b» towards the point «a». The highest value w_i will be obtained by the pair of values of the series which correspond to the point «a», and the lowest one – to the point «b».

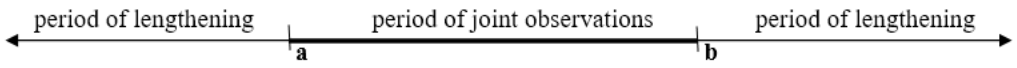

Figure 1. Location of the periods of lengthening

Thus, the closer some compared pair of values is to the period of restoration, the more they weigh when the distance value is calculated. If direction is not specified, then w_i is determined as it is equal to 1, and the formula looks like the Euclidean distance.

A new distance measure Time Series Distance (henceforward TSD) is proposed, which can be calculated on the basis of the similarity coefficient and Spearman's correlation coefficient, i.e. takes into consideration both the series similarity by their values and their relationship.

The following formula is used to calculate the distance:

$$d = \frac{2 - r_i - K_i}{3}, \tag{10}$$

where r_i – Spearman's correlation coefficient; K_i – similarity coefficient.

The computation formula K_i for data series of the same length will look like Jaccard formula:

$$K_i = \frac{L}{2 \times N - L}, \tag{11}$$

where N – the series length; L – the number of pairs of similar elements of series Q and S, is proposed to calculate in the following way:

$$L = \begin{cases} L + 1, & if\ Q_i \times (1 - p) \le S_i \le Q_i \times (1 + p) \\ L + 0 & - else \end{cases}, \tag{12}$$

where $p \in [0;1]$, default value was specified $p = 0.025$.

Using the right and the left parts of the inequality, we have obtained a variable range which will depend on the series Q behavior.

On fig. 2 the data series values are shown on the curves, presented by daily indices of water levels of post 79424 (river Stokhid in a small town Lyubeshiv, Volyn region) for the period from 01.04.2012 to 30.06.2012, the upper and lower boundaries have been constructed as well which were determined according to the formula (12) with $p = 0.025$. Thus, the more values of the series under comparison will happen to be in the range between the upper and lower boundaries (fig. 2), the higher the value L will be, and accordingly the value K_i.

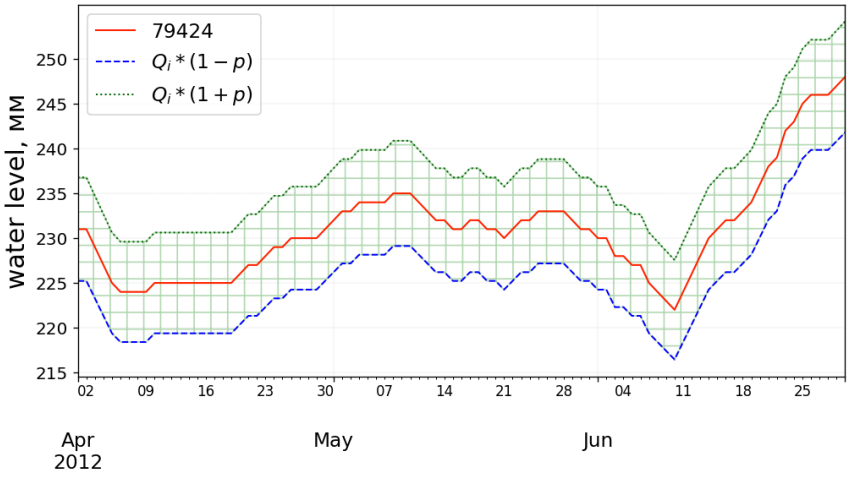

Figure 2. Curves of data series values of post 79424 and boundaries of permissible deviations

As the more general form, the formula (12) can be presented like:

$$L = \begin{Bmatrix} L+1, & if \;\; S_i \le (max(QS) - |min(QS)|) \times p \\ L+0 - else \end{Bmatrix} \quad (13)$$

where QS – is a set of data series taking part in the analysis, including the series Q.

The right side of the inequality in the formula (13) is determined taking into account the range of the whole set of data, so such presentation can be used for the distance matrix construction as well, for example, in data series clustering.

Taking into account, that $r \in [-1;1]$, and $K_i \in [0;1]$, due to the constant values of the numerator and the denominator in formula (10) we have obtained $d_{js} \in [0;1]$

The general scheme of the information technology of similar data series determination is presented on fig. 3.

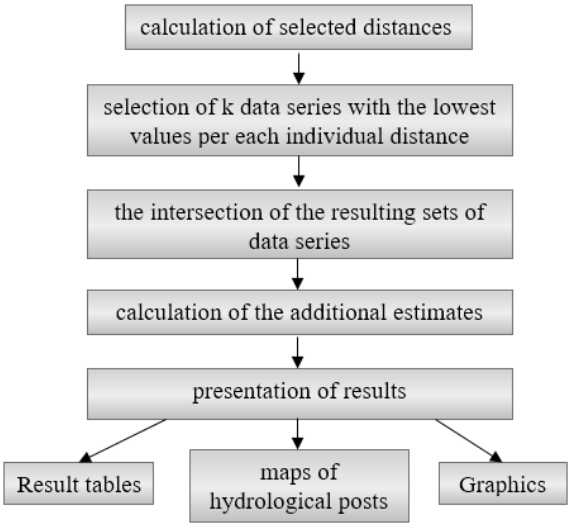

Figure 3. General scheme of the information technology

The selected distances are calculated on the first stage. As a result, the table of the following structure is formed: the names of the distances are located in the columns, the numbers of posts are on the lines, the distance values are on their intersection.

Having formed the table of the calculated distances, one should select those data series which have the lowest values according to all distances. In general, the selection procedure looks like:

1. The user defines the percentage of the total number of series, which are assumed to be the similar ones, thus, the value k is obtained.

2. By each distance d_i the first objects k are selected with the lowest values of the distance, the others are not taken into consideration.

To determine a set of similar series, an intersection between the sets of objects left after conduction of the p. 2 of the above-mentioned procedure is defined.

MAPE(%) and MSE, as well as the correlation coefficient Pearson r are calculated on the stage of additional estimates calculation whose values enable us to assess additionally the similarity of the defined series and the results in general. The list of data of estimates is not limited.

The library methods pandas of the programming language python are used to reduce the data to average weekly and average monthly values. The object DataFrame have methods of working with the data having time marks. The series values are reduced to the average weekly ones by using the following method:data.resample(freq).mean(), where data – pandas. DataFrame, containing the correspondent data series; resample() – the method provides grouping based on time marks; freq – of string type, determines the grouping density. The values used in the paper: W – to

obtain average weekly values and M – for average monthly ones; mean() – aggregation method, in this case, average values selection for the period according to the parameter freq.

Due to the results of the technology implementation, not only the final results was provided but some intermediate data like tables, curves, maps were presented as well.

Problem setting and the output data of the experiment. To conduct a computation experiment, the series of hydrological indices were chosen, presented by daily values of water levels fixed by 94 hydrological posts located on the water objects in the river Dnieper basin, Ukraine. The hydrological data were obtained from the Dnieper regional center of hydrometeorology.

One should determine the similar data series among other 93 hydrological posts for the series of values of the hydrological post 79545 located on the river Sluch of Novohrad-Volynskyi district, Zhytomyr region.

The period from 01.01.2010 to 31.12.2014 was chosen as a period of common observations when the similarity between the series will be determined.

We suppose, that the further lengthening of the data series of the post 79545 for the period from 01.01.2006 to 31.12.2009 will be conducted. The formula (8) or (9) is chosen depending on the choice of the lengthening period. We assume, that not more than 5% of the total number of series can be considered as similar ones (we round off up), i.e. the parameter $k = 5$. The value of the parameter p for the formula (11) was determined as 0,025.

We have to perform calculations on different presentations of data series: daily, average weekly, average monthly indices of water levels. As the range of values of most calculated distances is unlimited, the obtained results are normalized by the division of the correspondent distance by the maximum value, that will enable to reduce all values to single range [0;1]. The value of the distance TSD shouldn't be normalized.

The conclusions should be made concerning the possibility of using the found similar series to lengthen the data series of the post 79545 on the basis of the analysis of the obtained results by different ways of data series presentation.

Results of the computation experiment. Due to the results of conducted calculation experiment, first of all, the map of Ukraine was formed with the marked sites of hydrological posts located on it (fig. 4), whose series of values took part in the analysis. The hydrological post of data series, for which the similar series search was conducted, is marked by a circle (post №79545 on the territory of Zhytomyr region), and other posts in the Dnieper basin, whose series of values took part in the analysis, are marked by quadrangles.

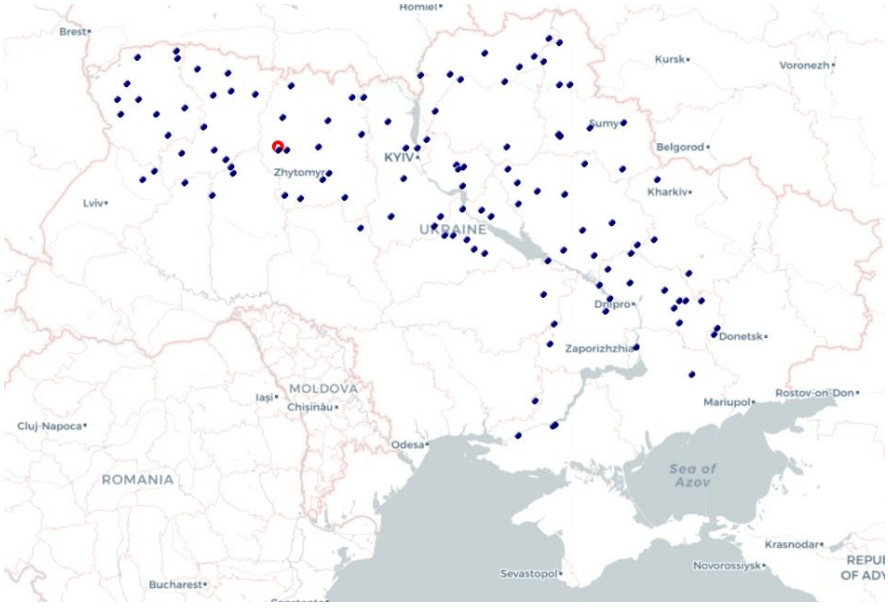

Figure 4. Map of Ukraine. Location of hydrological posts in the Dnieper basin

The daily water level values of the post and 93 data series, having common observation period with the series of the post 79545, are shown on fig. 5.

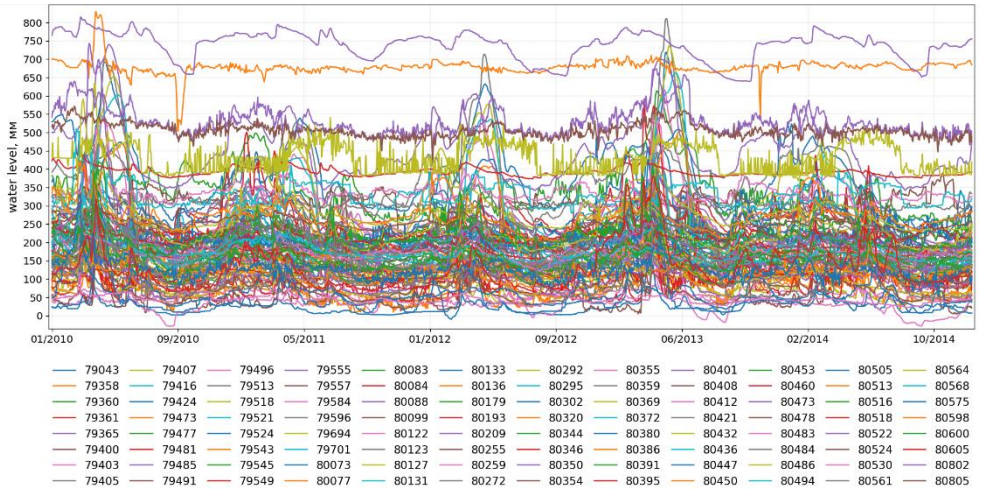

Figure 5. Graph of daily water level values for the period from 01.01.2010 to 31.12.2014

The normalized values of calculated distances for the series specified by the program as a result of calcualtions are presented in table 1.

Table 1

Normalized values of distances

Values of the series	Post	Distances					
		Euclidian	Euclidian weight	Manhattan	DTW	TSD	LCS_{dist}
Daily	79555	0,0308	0,0327	0,0224	0,0098	0,315	0,4803
	79694	0,0328	0,0355	0,0232	0,0114	0,334	0,5175
Average weekly	79555	0,0245	0,0273	0,0186	0,0152	0,315	0,5667
	79694	0,0271	0,0297	0,0215	0,0166	0,330	0,5833
	79596	0,0358	0,0366	0,0285	0,0195	0,392	0,5500
Average monthly	79555	0,0289	0,0312	0,0217	0,0141	0,288	0,5096
	79694	0,0313	0,0340	0,0227	0,0148	0,330	0,5441
	79596	0,0412	0,0419	0,0308	0,0172	0,377	0,5287

The results of calculations of estimates оцінок MSE and MAPE, as well as the correlation coefficient r for the data series, selected by the calculated distances, and the series of the post 79545 are presented in table 2.

Table 2

The results of the calculation of estimates

Representation of the values of the series	Post	Estimates		
		MSE	MAPE,%	r
Daily	79555	322,4	7,85	0,88
	79694	365,1	8,13	0,86
Average weekly	79555	283,2	7,67	0,89
	79694	332,9	8,00	0,86
	79596	577,5	10,93	0,82
Average monthly	79555	203,3	6,72	0,90
	79694	249,7	7,71	0,88
	79596	436,3	10,19	0,84

A fragment of the map of Ukraine with the sites of hydrological post 79545 location and hydrological posts whose data series are defined as similar ones to the data series of the post 79545 is presented on fig. 6.

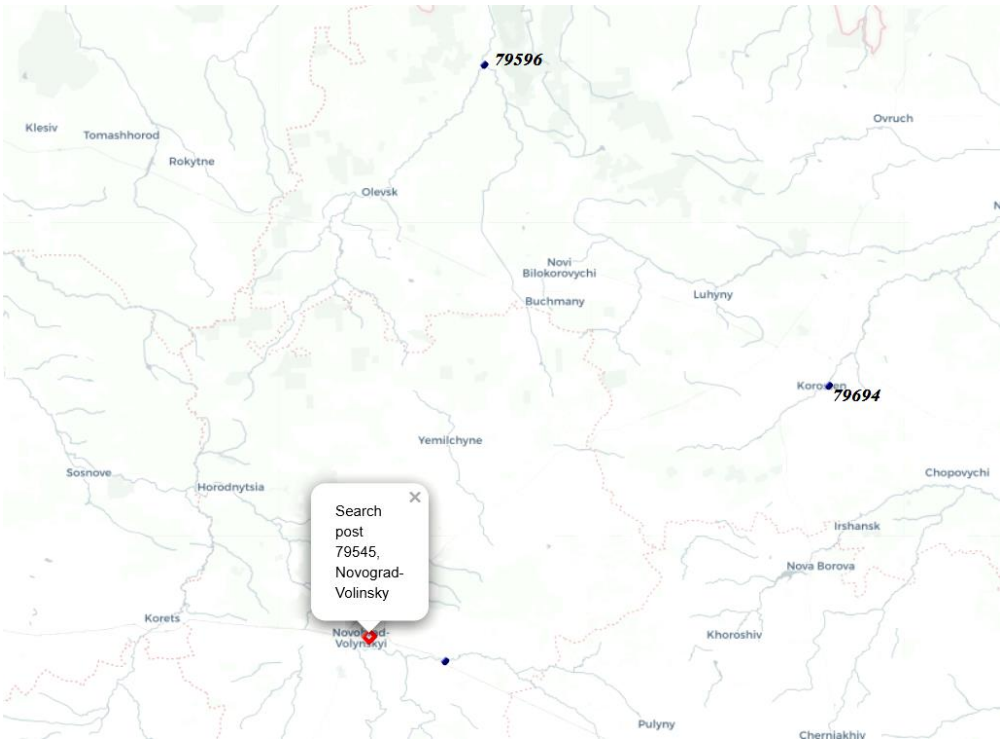

Figure 6. A fragment of the map of Ukraine. Location of hydrological stations

The values of data series indices of hydrological post 79545 and the posts which were defined as similar ones, are shown on fig. 7 according to the calculations results on the values reduced to daily water level ones

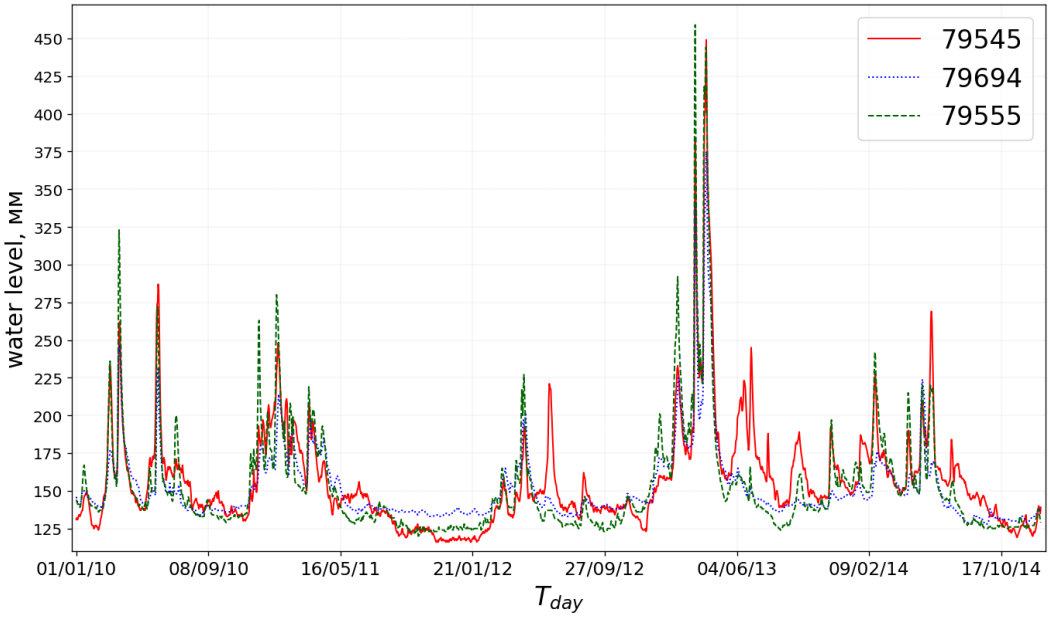

Figure 7. Graph of daily water level values as a result of the similar posts search

The values of data series indices of hydrological post 79545 and the posts which were defined as similar ones, are shown on fig. 8-9 according to the calculations results on the values reduced to weekly and monthly water level ones.

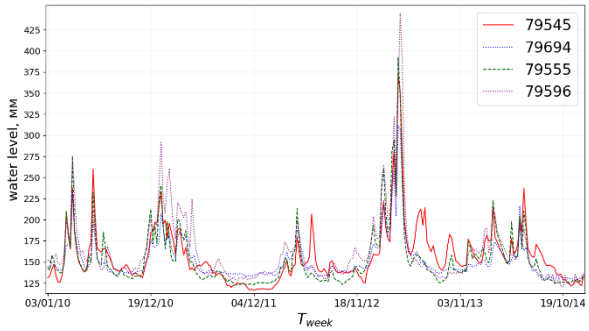

Figure 8. Graph of average weekly water level values as a result of the similar posts search

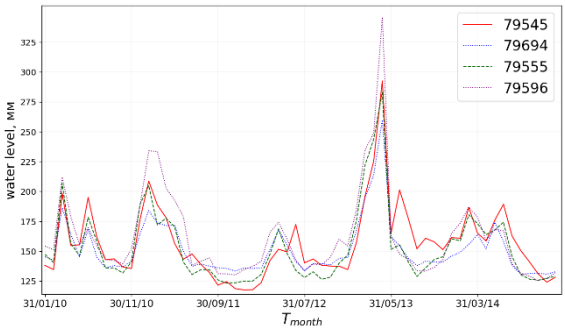

Figure 9. Graph of average monthly water level values as a result of the similar posts search

Discussion of the results. Three the most similar series of the water level indices were selected from 93 series, which were used for the comparison. Among them: a number of indices of hydrological post 79555, located on the river Tnya in the village Bronyky of Novohrad-Volynskyi district, Zhytomyr region; a number of indices of hydrological post 79596, located on the river Ubort in the village Perha Olevsk district, Zhytomyr region; a number of indices of hydrological post 79694, located on the river Uzh in the town Korosten, Zhytomyr region.

Thus, of all posts, whose location was presented on fig.4, and, as a result of the analysis, the selected hydrological posts were located on the territory of the same region (fig. 6).

Comparing the fig. 4, where all the series of water level indices are presented, with fig. 7–9, one can see, how close the selected series are to the series of post 79545 indices.

The series of water level indices values of posts 79555 and 79694 are the most similar, i.e. those which have the lowest values of the calculated distances to the series of indices values of hydrological post 79545, irrespectively of the way of indices values presentation (daily, weekly, monthly).

Having paid attention to the results, presented in table 2, one can see, that the selected data series of posts 79555 and 79694 have the lowest values of estimation MSE. MAPE estimates show the value of the difference between the values series of the post 79454 and the posts 79555 and 79694, expressed as a percentage, and they are less than 8,5% irrespectively to their daily, average weekly, or average monthly values, that has proved the similarity of the above-mentioned series. Besides, the correlation coefficient value for the data series of posts 79555 and 79694 with the post 79545 is higher than 0,85, which has proved the available direct relationship between the values of water level indices for the specified series.

As for the obtained results concerning the series of data of the hydrological post 79596, it should be mentioned, that this series was defined as a similar one by the technology. Nevertheless, according to the presented results in the table 1–2 one can make the conclusion, that it is necessary to carry out some extra study concerning the fact if the available series increase or decrease the quality of the further analysis when the specified series is included in it. Some doubts about using the data series of the hydrological post 79596 are caused by fact that for this series the value of estimation MSE calculated for average weekly average monthly values is 1,5 times higher than the same values of data series of hydrological posts 79555 and 79694. The doubts are proved by the fact that due to the calculations, based on data series presented by daily indices, the technology has removed the above-mentioned series from the list of similar series.

Conclusion. The developed and implemented technology has enabled us to find similar series on the basis of the calculated distance values. Some separate sets of series, defined as similar ones, were obtained for each distance that had been used. The intersection taken between the specified sets and the obtained single set of series provide a more stable solution while similar series are determined. An example of up to five similar series found is taken into consideration in the paper under discussion. According to the results of calculations on the average weekly and the average monthly indices, only three such series have been found by the technology, which belong to hydrological posts 79555,79694, 79596, and two series – while the calculations were done on the daily values of water level of the posts 79555 and 79694.

The obtained results have proved, that some other series were included into every five selected values by every separate distance, which were not taken into account when the intersection between the sets was considered.

The scientific novelty: the weighted Euclidean distance was described and used taking into account the valid data; some new distance measure was described and used enabling us to take into account both the degree of similarity between the series values and their correlation ties as well; the technology of similar series determination by the set of selected distances was developed. The practical value of the developed and implemented technology contributes to the following possibilities: it can be used for the data series which can be presented by the indices of different applied fields; making assessment and similar series determining, the intermediate stage of analysis, in particular; moreover, the proposed distance measures allow to increase the quality of similar series or their grouping determination. Some further research is intended to study the possibilities of data series lengthening and the gaps filling with the indices of other series defined as similar ones.

References

1. Liao T. W., Clustering of time series data – A survey, Pattern Recognit. Vol. 38. No. 11. Nov. 2005. P. 1857–1874. DOI: https://doi.org/10.1016/j.patcog.2005.01.025
2. Saxena A., et. al. A review of clustering techniques and developments. Neurocomputing, 267, 2017. P. 664-681. DOI: https://doi.org/10.1016/j.neucom.2017.06.053
3. Zhu X., Li Y., Wang J., Zheng T., Fu J. Automatic Recommendation of a Distance Measure for Clustering Algorithms. ACM Transactions on Knowledge Discovery from Data (TKDD), 15 (1). 2020. P. 1–22. DOI: https://doi.org/10.1007/978-81-322-1665-0_17
4. Savchuk T. O. Viznachennya evklidovoyi vidstani mizh nadzvichaynimi situatsiyami na zaliznichnomu transporti pid chas klasternogo analizu, Naukovi pratsi Vinnitskogo natsionalnogo tehnichnogo universitetu. – Seriya "Informatsiyni tehnologiyi ta komp'yuterna tehnika". 2010. No. 3. 2010.
5. Keogh E. J., Pazzani M. J. Derivative dynamic time warping. In Proceedings of the 2001 SIAM international conference on data mining. Society for Industrial and Applied Mathematics. 2001. April. P. 1–11. DOI: https://doi.org/10.1137/1.9781611972719.1
6. Dau H. A., Silva D. F., Petitjean F. et al. Optimizing dynamic time warping's window width for time series data mining applications. Data Mining and Knowledge Discovery 32. 2018. P. 1074–1120. DOI: https://doi.org/10.1007/s10618-018-0565-y
7. Raida V., Svoboda P., Rupp M. Modified dynamic time warping with a reference path for alignment of repeated drive-tests. In 2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall) IEEE. 2020. P. 1–6. DOI: https://doi.org/10.1109/VTC2020-Fall49728.2020.9348487
8. Senin P. Dynamic time warping algorithm review. Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA, 2008, 23 p.
9. Kate R. J. Using dynamic time warping distances as features for improved time series classification. Data Mining and Knowledge Discovery, 30 (2). 2016. P. 283–312. DOI: https://doi.org/10.1007/s10618-015-0418-x
10. Hu Z., Mashtalir S. V., Tyshchenko O. K., Stolbovyi M. I. Clustering matrix sequences based on the iterative dynamic time deformation procedure. International Journal of Intelligent Systems and Applications,10 (7). 2018. P. 66–73. DOI: https://doi.org/10.5815/ijisa.2018.07.07
11. Hunt J.W., Szymanski T. G. A fast algorithm for computing longest common subsequences. Communications of the ACM. Vol. 20. No. 5. 1977. P. 350–353. DOI: https://doi.org/10.1145/359581.359603
12. Hirschberg, Daniel S. Algorithms for the longest common subsequence problem. Journal of the ACM (JACM) 24.4. 1977. P. 664–675. DOI: https://doi.org/10.1145/322033.322044

13. Wan, Qingguo, et al. A fast heuristic search algorithm for finding the longest common subsequence of multiple strings. Twenty-Fourth AAAI Conference on Artificial Intelligence. 2010. P. 1287–1292. DOI: https://doi.org/10.1609/aaai.v24i1.7493
14. Wang Q., Dmitry K., Shang Y. Efficient dominant point algorithms for the multiple longest common subsequence (MLCS) problem. Twenty-First International Joint Conference on Artificial Intelligence. 2009. P.1494–1499.
15. Korkin D., Wang Q. Shang Y. An efficient parallel algorithm for the multiple longest common subsequence (MLCS) problem. 37th International Conference on Parallel Processing. IEEE, 2008. P. 354–363. DOI: https://doi.org/10.1109/ICPP.2008.79
16. Wang X., Mueen A., Ding H., Trajcevski G., Scheuermann P., Keogh E. Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery, 26 (2). 2013. P. 275–309. DOI: https://doi.org/10.1007/s10618-012-0250-5
17. Hryhorovych V. Analiz metryk dlia intelektualnykh informatsiinykh system, Visnyk Natsionalnoho universytetu "Lvivska politekhnika" "Informatsiini systemy ta merezhi". 2021. 9. P. 96–111. URL: https://doi.org/10.23939/sisn2021.09.096
18. Baturinets A., Antonenko S. Longest common subsewuence in the problem of determining the similarity of hydrological data series, Deutsche Internationale Zeitschrift für zeitgenössische Wissenschaft. 2021. No. 18. P. 62–64.

Список використаної літератури

1. Liao T. W. Clustering of time series data – A survey. Pattern Recognit. Vol. 38. No. 11. Nov. 2005. P. 1857–1874. DOI: https://doi.org/10.1016/j.patcog.2005.01.025
2. Saxena A., Prasad M., Gupta A., Bharill N., et. al. A review of clustering techniques and developments. Neurocomputing. 267. 2017. P. 664–681. DOI: https://doi.org/10.1016/j.neucom.2017.06.053
3. Zhu X., Li Y., Wang J., Zheng T., Fu J. Automatic Recommendation of a Distance Measure for Clustering Algorithms. ACM Transactions on Knowledge Discovery from Data (TKDD), 15 (1). 2020. P. 1–22. DOI: https://doi.org/10.1007/978-81-322-1665-0_17
4. Савчук Т. О., Петришин С. І. Визначення евклідової відстані між надзвичайними ситуаціями на залізничному транспорті під час кластерного аналізу. Наукові праці Вінницького національного технічного університету. Серія «Інформаційні технології та комп'ютерна техніка». 2010. Випуск № 3. 2010. 8 с.
5. Keogh E. J., Pazzani M. J. Derivative dynamic time warping. In Proceedings of the 2001 SIAM international conference on data mining. Society for Industrial and Applied Mathematics. April 2001. P. 1–11. DOI: https://doi.org/10.1137/1.9781611972719.1
6. Dau H. A., Silva D. F., Petitjean F. et al. Optimizing dynamic time warping's window width for time series data mining applications. Data Mining and Knowledge Discovery 32. 2018. P. 1074–1120. DOI: https://doi.org/10.1007/s10618-018-0565-y
7. Raida V., Svoboda P., Rupp M. Modified dynamic time warping with a reference path for alignment of repeated drive-tests. In 2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall) IEEE. 2020. P. 1–6. DOI: https://doi.org/10.1109/VTC2020-Fall49728.2020.9348487
8. Senin P. Dynamic time warping algorithm review. Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA, 2008, 23 p.
9. Kate R. J. Using dynamic time warping distances as features for improved time series classification. Data Mining and Knowledge Discovery, 30 (2). 2016. P. 283–312. Doi:10.1007/s10618-015-0418-x.
10. Hu Z., Mashtalir S. V., Tyshchenko O. K., Stolbovyi M. I. Clustering matrix sequences based on the iterative dynamic time deformation procedure. International Journal of Intelligent Systems and Applications,10 (7). 2018. P. 66–73. DOI: https://doi.org/10.5815/ijisa.2018.07.07
11. Hunt J.W., Szymanski T. G. A fast algorithm for computing longest common subsequences. Communications of the ACM. Vol. 20. No. 5. 1977. P. 350–353. DOI: https://doi.org/10.1145/359581.359603
12. Hirschberg, Daniel S. Algorithms for the longest common subsequence problem. Journal of the ACM (JACM) 24.4. 1977. P. 664–675. DOI: https://doi.org/10.1145/322033.322044
13. Wan, Qingguo, et al. A fast heuristic search algorithm for finding the longest common subsequence of multiple strings. Twenty-Fourth AAAI Conference on Artificial Intelligence. 2010. P. 1287–1292. DOI: https://doi.org/10.1609/aaai.v24i1.7493
14. Wang Q., Dmitry K., Shang Y. Efficient dominant point algorithms for the multiple longest common subsequence (MLCS) problem. Twenty-First International Joint Conference on Artificial Intelligence. 2009. P.1494–1499.
15. Korkin D., Wang Q. Shang Y. An efficient parallel algorithm for the multiple longest common subsequence (MLCS) problem. 37th International Conference on Parallel Processing. IEEE, 2008. P. 354–363. DOI: https://doi.org/10.1109/ICPP.2008.79

16. Wang X., Mueen A., Ding H., Trajcevski G., Scheuermann P., Keogh E. Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery, 26 (2). 2013. P. 275–309. DOI: https://doi.org/10.1007/s10618-012-0250-5
17. Григорович В. Аналіз метрик для інтелектуальних інформаційних систем. Вісник Національного університету «Львівська політехніка». «Інформаційні системи та мережі». 2021. Вип. 9. С. 96–111. URL: https://doi.org/10.23939/sisn2021.09.096
18. Батурінець А., Антоненко С. Найдовша спільна підпослідовність в задачі визначення схожості гідрологічних рядів даних. Deutsche Internationale Zeitschrift für zeitgenössische Wissenschaft. 2021. № 18. С. 62–64.

УДК: 004.67:519.25

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ВИЗНАЧЕННЯ СХОЖИХ РЯДІВ ДАНИХ ІЗ ВИКОРИСТАННЯМ МІР ВІДСТАНЕЙ

Анастасія Батурінець

Дніпровський національний університет імені Олеся Гончара, Дніпро, Україна

Резюме. *Метою є розроблення та реалізація технології визначення схожих рядів даних, а також її апробація на рядах даних, представлених гідрологічними показниками. Предметом дослідження є методи та підходи визначення схожих рядів даних. Об'єктом дослідження є процес визначення схожих рядів даних, представлених певними показниками. Завдання: запропонувати й реалізувати міри відстані, одна з яких враховує схожість між значеннями рядів даних та їх зв'язок, а друга – заснована на зваженій евклідовій відстані, але з урахуванням необхідності актуалізації даних, які є важливішими за певних умов задачі. Реалізувати технологію визначення схожих рядів даних, представлених певними показниками. Для стійкішого розв'язку реалізувати процедуру визначення набору схожих рядів на підставі отриманих результатів за кожною окремою відстанню. Проаналізувати отримані результати та зробити висновки щодо можливості практичного використання технології. Використовуваними методами є методи статистичного аналізу, методи обчислення відстаней та схожості між рядами. Отримані результати реалізовано технологію визначення схожих рядів даних. Як складову технології реалізовано дві запропоновані й описані міри відстаней. Реалізовано процедуру визначення набору схожих рядів за отриманими значеннями відстаней. Наукова новизна: описано та застосовано евклідову зважену відстань з урахуванням актуальності даних. Описано та застосовано нову міру відстані, яка дозволяє врахувати як ступінь подібності між значеннями рядів, так і їх кореляційний зв'язок. Розроблено технологію визначення схожих рядів за множиною обраних відстаней. Практична значущість розробленої та реалізованої технології полягає в таких можливостях: застосування на рядах даних різних прикладних областей; проведення оцінювання та визначення схожих рядів, зокрема як проміжний етап аналізу. Крім того, запропоновані міри відстані дозволяють підвищити якість визначення схожих рядів або їх групування. Подальші дослідження планується спрямувати на дослідження можливостей подовження рядів даних та поповнення пропусків за значеннями показників інших рядів, визначених як схожі.*

Ключові слова: *міри відстані, схожість числових рядів, LCS, DTW, TSD, подібність рядів даних, гідрологія.*

https://doi.org/10.33108/visnyk_tntu2022.01.128 *Отримано 18.01.2022*