ORIGINAL RESEARCH
published: 09 September 2021
doi: 10.3389/fphy.2021.714910

Similarity Analysis of Alarm Sequences by a Shuffling Method

Yifan Lin[1], Shengfeng Wang[2]*, Ye Wu[3] and Jinghua Xiao[1]*

[1]School of Science, Beijing University of Posts and Telecommunications, Beijing, China, [2]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China, [3]School of Journalism and Communication, Beijing Normal University, Beijing, China

Modern telecommunication systems produce large amounts of alarm messages, and alarm management is vital for telecommunication systems' high-quality performance. Building functional networks by observing the pair similarity between time series is a useful way to filter and reduce alarm messages. Because of the coexistence of positive and negative correlations among telecommunication devices, most of the similarity measures have troubles in computing the complex correlations. In this paper, we propose an index of measuring how much two-alarm series deviate from the uncorrelated situation to detect the correlation of both sides. Synthetic sequences verify our method. Furthermore, we apply our method to analyze telecommunication devices' alarm correlation in a province of China. Our index of pair similarities is capable of measuring other discrete event data.

Keywords: telecommunication alarm, similarity measure, alarm correlation network, event relationship network, time series

OPEN ACCESS

Edited by:
Gaogao Dong,
Jiangsu University, China

Reviewed by:
Jianguo Liu,
Shanghai University of Finance and Economics, China
Xiao-Pu Han,
Hangzhou Normal University, China
André L. M. Vilela,
Universidade de Pernambuco, Brazil

Correspondence:
Shengfeng Wang
sfwang@bupt.edu.cn
Jinghua Xiao
jhxiao@bupt.edu.cn

Specialty section:
This article was submitted to Social Physics, a section of the journal Frontiers in Physics

Received: *26 May 2021*
Accepted: *19 July 2021*
Published: *09 September 2021*

Citation:
Lin Y, Wang S, Wu Y and Xiao J (2021) Similarity Analysis of Alarm Sequences by a Shuffling Method. Front. Phys. 9:714910. doi: 10.3389/fphy.2021.714910

INTRODUCTION

According to the Ministry of Industry and Information Technology of China, the total number of mobile phone users reached 1.594 billion and more than 98 percent of administrative villages had access to optical fiber and 4G in China at the end of 2020. Numerous base stations and other kinds of equipment constitute huge telecommunication networks with complicated structures. These telecommunication systems produce a large number of alarm messages every day, which pose a challenge to faults management. In the course of the managing process, various telecommunication devices may affect each other [1, 2]. To effectively manage the system, it is critical to develop strategies for correlating alarm messages by the physical connections of network elements or knowledge derived from alarm experiences.

To perform fault management under a large number of alarm messages, it is important to reduce the number of alarm messages by correlating different devices' messages. In telecommunication networks, some expert systems are implemented to filter and correlate alarms. Italy [3], first uses expert system rules to recognize alarm correlation patterns and instantiate network fault hypotheses, and then applies a heuristic search to determine the best solution among the hypotheses. ALLINK™ Operations Coordinator from NYNEX [4] uses an expert system to filter network alarms. Most of the existing expert systems are for relating fault messages, and transferring the knowledge of human experts into an automated system. Other related methodologies were proposed. The work in [5, 6] is based on a formal language representation of the communication system. A. Bouloutas in [5] focuses on identifying errors in a known protocol: it is not an alarm correlation as such. The problem considered by A. Bouloutas and S. Calo in [6] is fault localization from alarms. It is a related although different problem. Such researches do not consider the occurring time of alarms, and assume knowledge of the network topology.

With the continuous development of telecommunication systems, telecommunication networks are becoming more complex, and features such as heterogeneous devices, network structures, and technologies are coexisting and cooperating within the system. This is a problem when domain-experts build management systems for root cause analysis or event relationship networks (ERNs). Data-driven fault management may be helpful [7]. Perng [8] utilized the event history logs in shorting the ERNs design process and perfecting the quality of ERNs. Besides constructing the ERNs, one can build device-device correlation networks from alarm logs. Particularly, telecommunication devices are deployed over large geographical area, and the device-to-device networks could be useful in understanding the performance of the whole systems. Based on the discrete alarm time of devices, device-to-device networks can be constructed by correlating alarm series to form a functional network. Functional structures are of great importance in aiding understanding of the properties of various man-made and natural networks [9, 10]. Differing from physical structures, functional structures are generally built by observing the similarity between time series. Depending on the application scenario and the type of data, there are various way of computing pair similarity. Euclidean Distance is the most basic measure, when two sequences are of equal length. To measure the similarity of unequal-length time series, Dynamic Time Warping (DTW) [11] is useful, and is used in many proposed optimizations [12–14]. If two time series have similar morphology in most time periods but only have certain differences in a very short time, Euclidean Distance and DTW cannot accurately measure the similarity between them, which can be solved by Longest Common Subsequence (LCSS) [15]. However, the measures mentioned above only focus on calculating how different the two series are, and ignore the probability of them being such different. Furthermore, due to the complexity of the system, recovering alarm messages sometimes needs to check both positively and negatively correlated devices. Most of the similarity measures may encounter troubles here.

To tackle the problem above we propose an index built on measuring to what extent the two series deviate from the corresponding shuffled series, to score the pair similarity. We then construct synthetic series to verify the method. Furthermore, we apply our method to analyze the alarm correlation of telecommunication devices in a province of China. Although our method focuses on the application of the telecommunication devices' alarm series, it can also be applied to general discrete event data.

METHODS

The Definition of the Similarity Score Between Alarm Sequences

This section defines an index to score the similarity.

Firstly, let S_i be the time sequence representing the alarm timing of ith device.

$$S_i = \left\{s_1^i,\ s_2^i,\ \cdots,\ s_k^i,\ \cdots,\ s_{|S_i|}^i\right\}, \tag{1}$$

where $|S_i|$ denotes the size of S_i. Given two sequences S_A, S_B and $n = |S_A \cap S_B|$ being the number of the same timestamps between them, we calculate the possibility of two sequences still having n same timestamps when they are randomly shuffled. In detail, let $|S_A| = m_1$, $|S_B| = m_2$, and the total time duration is assumed to be D seconds. This may be illustrated by comparing our method to a textbook example in probability. In this model, D balls are numbered and put in an opaque box. Person A first picked out m_1 balls randomly and put them back after recording their numbers. Then, person B picked m_2 balls out at random and recorded the number too. The probability of n balls being picked out twice can be expressed as

$$P_n = \frac{C_{m_1}^{n} C_{D-m_1}^{m_2-n}}{C_D^{m_2}},\ 0 \le n \le \min\{m_1,\ m_2\}, \tag{2}$$

where $C_D^{m_2}$ is the number of possible combinations of m_2 balls that person B can pick out from the box, and $C_{m_1}^{n}$ and $C_{D-m_1}^{m_2-n}$ compute the number of possible combinations in which B picks n balls in common with m_1 recorded balls and $m_2 - n$ from the $D - m_1$ unrecorded balls respectively. It is obvious that $\sum_{n=0}^{\min\{m_1,\ m_2\}} P_n = 1$. **Eq. 2** computes the probability of having n timestamps in common between S_A and S_B when they are randomly shuffled.

We define the similarity index in terms of P_n. According toP_n's definition, its value would be no less than 0 and no more than 1. If a small P_n, such as less than 0.05, appears, it means that a rare event has occurred, which results from the appearance of a much larger or smaller n compared to its expectation of two uncorrelated sequences. When n is much larger than the expectation, it shows that two devices send alarm messages together more often than the random case, and vice versa. A large n means that one devices alarm may be caused by an alarm in the other, while a small n may mean that one devices alarm is caused by the normal function of the other. Both cases leads to the conclusion that devices A and B are correlated. Because a large P_n represents that the correlation between A and B has no difference from the random case, we define the index which scores the correlation of alarm sequences A and B as

$$c_{AB} = 1 - P_n, \tag{3}$$

which is symmetric, so that $c_{AB} = c_{BA}$.

Computational Processing

Large values of D, such as in several days of data, would make the similarity computation very expensive. Therefore we separated the total time duration into several windows with equal size and computed the c_{ij} of two devices within each window respectively. Then, the average value $\bar{c}_{ij}$ over each window is taken as the final similarity score that describes the degree of correlation between two devices.

Equation 2 is the probability mass function (pmf) of hypergeometric distribution. When the total seconds D is a large number, for instance, more than 10,000, it is hard to calculate the value of $C_D^{m_2}$ because the factorial of D is too large for computer to store as m_2 increases. Here, when both the value of m_1 and m_2 are more than 90, we use an

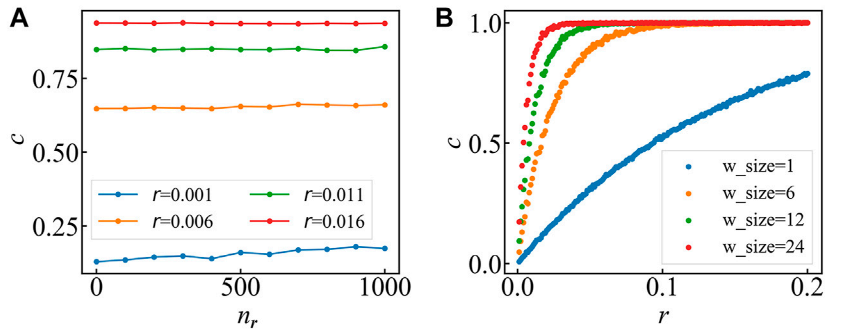

FIGURE 1 | (A) Average similarity scores at different correlation level. The sequences of device A and B are generated 10 times. The results showed in this figure are the average of 10 experiments when the window size is 24. Similarity score barely changes as n_r increases. The line moves upwards when actual correlation r increases. **(B)** Average similarity scores at different window size. The results are also the average of 10 experiments but when adding 1,000 random alarm timestamps into S_B. The range that probability score varies with r widens as the window size decreases.

approximation method proposed by Irving W. Burr [16], who found the approximation relation between the hypergeometric and Poisson distributions as

$$h(x; N, n, k) = p\left(x; \frac{kn}{N}\right)\left\{1 + \left(\frac{1}{2k} + \frac{1}{2n}\right)\left[x - \left(x - \frac{nk}{N}\right)^2\right] + O\left(\frac{1}{k^2} + \frac{1}{n^2}\right)\right\} \tag{4}$$

where $h(x; N, n, k)$ denotes hypergeometric probability for x in n given k in N and $p(x; kn/N)$ denotes Poisson probability with parameter kn/N. Thus, we use the adjustive factor below to approximate the hypergeometric probability by the Poisson probability when m_1 and m_2 are quite large.

The Verification of the Method

For verifying the validity of our method, we generated synthetic alarm series whose correlation can be set manually. Let $S_A = \{s_1^A, s_2^A, \cdots, s_k^A, \cdots, s_{|S_A|}^A\}$ be the series of device A which records 10,000 alarms within 58 days. Assuming that the probability of device B reporting an alarm message when device A reports is r, indicating the actual alarm correlation between two devices. We also add n_r random timestamps, which represent the random alarming of device B itself or correlating with other devices, into S_B to see if the score calculated by the method changes as the number of timestamps in S_B changes. **Figure 1A** shows the tendency of average similarity score with n_r under different correlation levels in 10 experiments. In the figure, similarity score barely changes as n_r increases, which means the proposed method is robust when the number of alarms changes. In addition, **Figure 1A** shows that the method is capable of distinguishing different r levels as the line moves upwards when actual correlation r increases.

We also study the influence of window size on the similarity scores. With random timestamps n_r fixed at 1,000, we calculated similarity scores at different r levels using the window size of 1 h, 6 h, 12 h and 24 h. In **Figure 1B**, when using window size of 24 h, the similarity score will be close to 1 if r is more than 0.04. It indicates that the method considers r being more than 0.04 as two sequences being strongly related. However, when we reducing the window size, the maximum value of the curve using the window size of 1 h is only near 0.8, meaning that the method could even distinguish r whose value is more than 0.2 which is not a small value when considering correlation between two devices. When analyzing real data, we can change the window size to make the distribution of scores as scattered as possible so that we can rank all the pairs of devices and find the most related ones.

EXPERIMENT

In this section, we use the method described above to analyze real data of device alarms in telecommunication networks and constructed a functional network that could help to locate faults by scoring the probability of every two devices being correlated when reporting alarm messages. Moreover, based on the location of devices, we construct a city-to-city alarm network (CCAN) and analyze its structure.

Data Description

The database is from a Chinese telecommunication company, including the alarm messages of about 500,000 telecommunication devices in a province of China from 26th August to 25th September in 2015. In the following, we anonymize the name of the province (named as G hereafter) and the related cities. Each message in the database includes device ID, alarm title, type, location, and other information. We pick out messages that recorded both device ID and location. **Figure 2** shows that the alarm number distribution of 508,636 devices follow a power law distribution. To obtain the main correlation structure, we preprocess the database and take devices that documented alarm messages between 600 and 50,000 times into account. After that, 6,527 devices are considered into the following analysis.

Result

Firstly, letting every device be the vertex, a fully connected network is formed. Here, an edge is equivalent to a pair of devices, and its weight equals the calculated similarity score. Then, we remove edges whose similarity scores are smaller than a threshold value, and the rest of the edges form the backbone of the alarm correlation network. Secondly, we use every device's

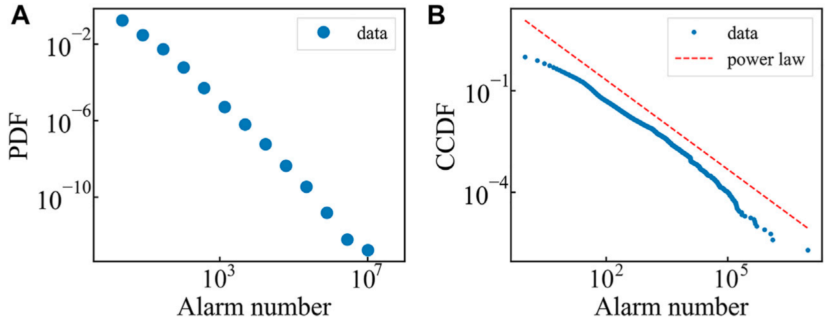

FIGURE 2 | (A) Probability density function (pdf) of alarm times with log-log coordinate. **(B)** Complementary cumulative distribution function (ccdf) of alarm times with log-log coordinate. The data follows a straight line in log-log coordinates, which indicates that the alarm number obeys the power law distribution.

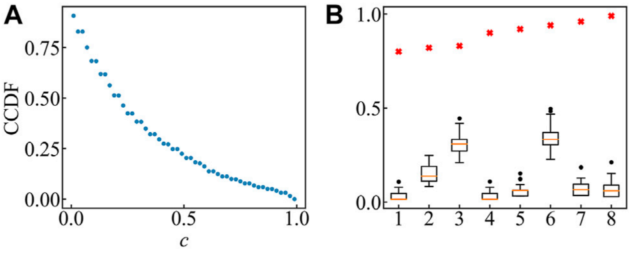

FIGURE 3 | (A) Complementary cumulative distribution function (ccdf) of the similarity scores. Each point shows the percentage of scores that are more than a certain value. When using a threshold that is no less than 0.8, more than 90% of the edges will be removed. **(B)** Boxplot of randomly shuffled alarm sequences. For every chosen pair of devices, shuffle their sequence for 100 times and calculate the similarity scores of them. Red crosses symbolize the original similarity score. The original scores are completely separated from the shuffled sequences scores' boxplots.

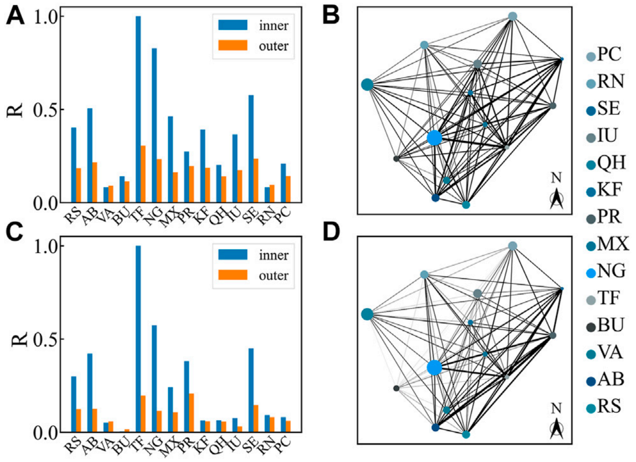

FIGURE 4 | Relative connection density and intercity alarm relevance network (CCAN): **(A)** Percentage of relative density inside and outside cities when using threshold 0.8. The inner relative density is painted blue and the outer is painted orange. Each *R* is normalized by the maximum value (0.1902) of inner relative density. **(B)** City-to-city network under threshold 0.8. Different cities are symbolized by dots with different colors. The size of dots represents the number of devices in every city. The width of edges is proportional to the value of relative density between cities. **(C)** Percentage of relative density inside and outside cities when using threshold 0.999. Each *R* is normalized by the maximum value (0.0664) of inner relative density. **(D)** City-to-city network under threshold 0.999. **(A,C)** show that when increasing the threshold of removing edges, the structure between cities starts emerging (still weaker than the connection inside cities). **(B,D)** show that cities lying in the southeast of the province are connected more strongly than otherwhere.

located city to analyze the relationship between the alarm numbers inside and outside the city and the network of cities.

After applying our method to the data, **Figure 3A** shows the complementary cumulative distribution function (CCDF) of the similarity scores. Each point shows the percentage of similarity scores that are more than a value. In **Figure 3A**, when we take 0.8 as a threshold of removing edges, there was less than 10% of the edges left in the network. In consideration of this, we use an approximation of hypergeometric distribution and the remaining term is of the same order as 0.0001, so choosing 0.999 as an upper bound for the analysis will not impact computation accuracy.

We randomly chose eight pairs of devices whose similarity scores are greater than 0.8 and shuffled their alarm times from the last 31 days to see if our index separates correlated devices from independent, uncorrelated ones. **Figure 3B** compares the scores of original alarm sequences with those of the shuffled sequences in 100 repetitions. The results show that the original scores are completely separated from the boxplots of the shuffled sequences scores, meaning that the device pairs left in the network are statistically correlated.

In the following, we show the devices' alarm correlation network in G province, China. To compare the connection strength inside and outside the cities, we normalize the connection by relative connection density. For every city in G province, the relative density inside the city is defined as

$$R_A^{in} = \frac{E_A}{|A| \times \frac{|A|-1}{2}}, \tag{5}$$

where E_A represents the number of edges (ignoring the similarity scores) and $|A|$ is the number of devices inside the city. The relative density outside the city is defined as

$$R_A^{out} = \frac{\sum_{B\in\mathcal{F},\, B\neq A} E_{AB}}{|A| \times \sum_{B\in\mathcal{F},\, B\neq A} |B|} \tag{6}$$

where $\mathcal{F}$ is a set of all the cities in province G, the numerator represents the number of edges between A and other cities, and the denominator represents the number of edges if all the devices inside city A are connected to all the devices inside other cities. We calculate the relative densities R after removing those whose similarity scores are less than 0.8, and present the percentage of the relative densities outside and inside the city in **Figure 4A**. Most of the cities' inner relative densities are more than the outer ones, which is consistent with our intuition. When increasing the threshold from 0.8 to 0.999, some structure between cities emerges, as shown by **Figure 4C**. Therefore, we draw the city-to-city network where the edges are weighted by the relative density between two cities which is defined as

$$R_{AB} = \frac{E_{AB}}{|A| \times |B|} \tag{7}$$

where E_{AB} is the number of edges between city A and B. The city-to-city networks under threshold 0.8 and 0.999 are exhibited in **Figures 4B,D**, where different cities are symbolized by colored dots, the size of dots represents the number of devices inside the city, and the width of the edges represents the relative densities between two cities. **Figure 4B** shows that the CCAN is a fully connected network. Devices from different cities connected more strongly than expected. Although the cities that lie in the north of G province, such as QH, RN and PC, have more devices than the cities in the southeast side, their connections with other cities are weaker than southeast cities. City NG and TF which are strongly connected to almost all cities in the province seem to be the center of the city-to-city network. However, when the threshold increases to 0.999, the center of the city-to-city network moves to the city PR whose device number is quite small when compared with other cities and PR is only connected strongly to the cities from the southeast. It seems that there are an alarm group consists of cities from the southeast area of the province.

CONCLUSION

Modern telecommunication systems produce large amounts of alarm messages. Correlating different alarm series in vital to effectively manage these alarm messages and maintain the performance of telecommunications networks. To measure the complex spatiotemporal correlation between telecommunication devices, we propose an index that uses the deviation of two alarm series from the random case to score the pair similarity in the device-to-device network. In **Figure 1**, synthetic series verify the validity of our index, and show that the similarity score can distinguish series pairs with different correlation levels and is robust when alarm numbers change. Moreover, the range that probability score vary with correlation level can be widened by reducing window size when calculating, as shown in **Figure 1B**. After verifying our method, we used it to analyze the telecommunication alarm database of devices in a Chinese province, and construct an alarm correlation network. In **Figure 4**, the results show that for most of the cities, the connection strength inside the cities is higher than outside. However, the connections outside cities are comparable with those inside cities. When increasing the edge removal threshold, cities' structures start to emerge (though still weaker than the connections within cities). By analyzing the CCAN, we find that cites lying in the southeast of the province connect more strongly than elsewhere. Our similarity score measures the pair similarity by deviating from the random case and has a potential for more general applications.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Because privacy issues are present, the data should not be shared. Requests to access these datasets should be directed to sfwang@bupt.edu.cn or linyifan13@foxmail.com.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and have approved it for publication.

FUNDING

This research was supported by the National Key R&D Program of China (grant No. 2020YFF0305300).

REFERENCES

1. Haes Alhelou H, Hamedani-Golshan M, Njenda T, and Siano P. A Survey on Power System Blackout and Cascading Events: Research Motivations and Challenges. *Energies* (2019) 12(4):682. doi:10.3390/en12040682
2. Chung H-M, Li W-T, Yuen C, Chung W-H, Zhang Y, and Wen C-K. Local Cyber-Physical Attack for Masking Line Outage and Topology Attack in Smart Grid. *IEEE Trans Smart Grid* (2019) 10(4):4577–88. doi:10.1109/TSG.2018.2865316
3. Brugnoni S, Bruno G, Manione R, Montariolo E, and Sisto L. An Expert System for Real Time Fault Diagnosis of the Italian Telecommunications Network. *Integrated Network Management III*. In: Proceedings of the IFIP TC6/WG66 Third International Symposium on Integrated Network Management with participation of the IEEE Communications Society CNOM and with support from the Institute for Educational Services; 18-23 April; San Francisco, California, USA. AE Amsterdam, Netherlands: North-Holland Publishing Co., Div. of Elsevier Science Publishers B.V (1993). p. 617–28.
4. Jakobson G, and Weissman M. Alarm Correlation. *IEEE network* (1993) 7(6): 52–9. doi:10.1109/65.244794
5. Bouloutas AT. *Modeling Fault Management in Communication Networks [Ph.D Thesis]*. Columbia: Columbia University (1990).
6. Bouloutas AT, Calo S, and Finkel A. Alarm Correlation and Fault Identification in Communication Networks. *IEEE Trans Commun* (1994) 42(234):523–33. doi:10.1109/TCOMM.1994.577079

7. Dai X, and Gao Z. From Model, Signal to Knowledge: A Data-Driven Perspective of Fault Detection and Diagnosis. *IEEE Trans Ind Inf* (2013) 9(4):2226–38. PubMed PMID: 13843496. doi:10.1109/TII.2013.2243743
8. Perng C-S, Thoenen D, Grabarnik G, Ma S, and Hellerstein J. Data-driven Validation, Completion and Construction of Event Relationship Networks. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 24 - 27; Washington, DC, USA. New York, NY, United States: Association for Computing Machinery (2003). p. 729–34. doi:10.1145/956750.956848
9. Albert R, and Barabási A-L. Statistical Mechanics of Complex Networks. *Rev Mod Phys* (2002) 74(1):47–97. doi:10.1103/RevModPhys.74.47
10. Newman MEJ. The Structure and Function of Complex Networks. *SIAM Rev* (2003) 45(2):167–256. Epub May 2, 2003. doi:10.1137/S003614450342480
11. Sakoe H, and Chiba S, editors. *A Dynamic Programming Approach to Continuous Speech Recognition*. Budapest, Hungary: International Congress on Acoustics (1971).
12. Sakoe H, and Chiba S. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. In: Waibel A and Lee K-F, editors. *Readings in Speech Recognition*. San Francisco: Morgan Kaufmann (1990). p. 159–65. doi:10.1016/b978-0-08-051584-7.50016-4
13. Keogh EJ, and Pazzani MJ. *Derivative Dynamic Time Warping*. Chicago: First SIAM international conference on data mining (2001).
14. Lahreche A, and Boucheham B. A Fast and Accurate Similarity Measure for Long Time Series Classification Based on Local Extrema and Dynamic Time Warping. *Expert Syst Appl* (2021) 168:114374. doi:10.1016/j.eswa.2020.114374
15. Golay X, Kollias S, Stoll G, Meier D, Valavanis A, and Boesiger P. A New Correlation-Based Fuzzy Logic Clustering Algorithm for FMRI. *Magn Reson Med* (1998) 40(2):249–60. doi:10.1002/mrm.1910400211
16. Burr IW. Some Approximate Relations between Terms of the Hypergeometric, Binomial and Poisson Distributions. *Commun Stat* (1973) 1(4):297–301. Epub 27 Jun 2007. doi:10.1080/03610927308827027

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.