Electronic Physician (ISSN: 2008-5842) http://www.ephysician.ir
June 2017, Volume: 9, Issue: 6, Pages: 4648-4654, DOI: http://dx.doi.org/10.19082/4648

# Providing an imputation algorithm for missing values of longitudinal data using Cuckoo search algorithm: A case study on cervical dystonia

Amin Golabpour[1], Kobra Etminani[2], Hassan Doosti[3], Hamid Heidarian Miri[4], Reza Ghanbari[5]

[1] M.Sc., Department of Biomedical Informatics, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran
[2] Ph.D., Assistant Professor, Department of Biomedical Informatics, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran
[3] Ph.D., Assistant Professor, Department of Biostatistics and Epidemiology, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran
[4] Ph.D., Assistant Professor, Management & Social Determinants of Health Research Center, Mashhad University of Medical Sciences, Mashhad, Iran
[5] Ph.D., Associate Professor, Faculty of Mathematical Sciences, Department of Applied Mathematics, Ferdowsi University of Mashhad, Mashhad, Iran

**Type of article:** Original

**Abstract**

**Background:** Missing values in data are found in a large number of studies in the field of medical sciences, especially longitudinal ones, in which repeated measurements are taken from each person during the study. In this regard, several statistical endeavors have been performed on the concepts, issues, and theoretical methods during the past few decades.

**Methods:** Herein, we focused on the missing data related to patients excluded from longitudinal studies. To this end, two statistical parameters of similarity and correlation coefficient were employed. In addition, metaheuristic algorithms were applied to achieve an optimal solution. The selected metaheuristic algorithm, which has a great search functionality, was the Cuckoo search algorithm.

**Results:** Profiles of subjects with cervical dystonia (CD) were used to evaluate the proposed model after applying missingness. It was concluded that the algorithm used in this study had a higher accuracy (98.48%), compared with similar approaches.

**Conclusion:** Concomitant use of similar parameters and correlation coefficients led to a significant increase in accuracy of missing data imputation.

**Keywords:** Missing data, Imputation of missing data, Cuckoo algorithm, Longitudinal data

## 1. Introduction

### *1.1. Background and objectives*

In longitudinal studies, the evaluation unit is subject to initial and repeated measurements over time. In such studies, missing values are inevitable because follow-up of some participants might be difficult due to lack of presence or unwillingness to cooperate. These missing data pose challenges in analyzing and modeling of data mining. A longitudinal survey is a correlational research study that involves repeated observations of the same variables over long periods of time, often many decades. It is often a type of observational study, although the study also can be structured as longitudinal randomized experiments (1). Three patterns of missingness are considered in longitudinal studies, structure of which is provided in Figure 1 (2):

1) Lack of access to patients at all measurement times (Figure 1.A)

**Corresponding author:**
Assistant Professor Dr. Kobra Etminani, Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.
Tel: +98.5138002442, Fax: +98.5138002445, Email: etminanik@mums.ac.ir
Received: November 29, 2016, Accepted: March 02, 2017, Published: June 2017
iThenticate screening: February 25, 2017, English editing: June 01, 2017, Quality control: June 10, 2017

2) Unwillingness to continue participation in the study (Figure 1.B)
3) A combination of the first and second items (Figure 1.C)

In the current study, we only focused on the missing data of part B, for which the proposed algorithm was considered. This multiobjective algorithm was implemented with the help of a Cuckoo optimization algorithm. Cervical Dustonia (CD) is described, and the structure of standard data for this disease is delineated. Thereafter, the missingness structure and Cuckoo search (CS) algorithm are outlined, the proposed algorithm is provided to estimate the missing data, and the algorithm is evaluated to ensure accuracy of data.

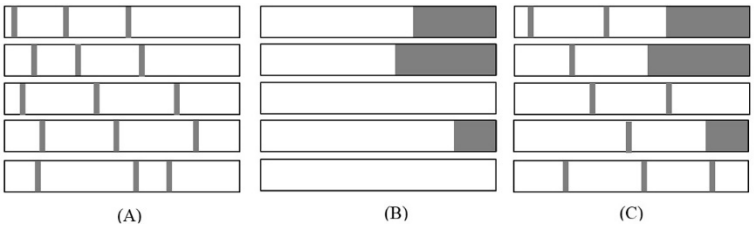


**Figure 1.** Missing data model in longitudinal studies (2)

### *1.2. Cervical dystonia*

Cervical dystonia, also called spasmodic torticollis, is a painful condition in which the neck muscles contract involuntarily, causing the head to twist or turn to one side. Cervical dystonia also can cause the head to uncontrollably tilt forward or backward (3). Cervical dystonia patient data included seven feature ID, week, site, treatment, age, sex, and total score of Toronto Western Spasmodic Torticollis Rating Scale (total TWSTRS score).

### *1.3. Types of missing data*

In general, three types of missing data mechanisms exist, which were developed by Little and Rubin (5). Understanding the reasons why data are missing is important to correctly handle the remaining data. If values are missing completely at random, the data sample is likely still representative of the population. There are three types of missing data: missing completely at random (MCAR), missing at random (MAR), missing not at random (MNAR).

### *1.4. Proposed algorithm*

At first, data preprocessing was conducted followed by the elimination of k% of data based on the missingness pattern B in Figure 1. It should be mentioned that the missingness mechanism of this K% is defined based on the number of rows. Afterward, the proposed algorithm was proposed.

### *1.5. Preprocessing*

Preprocessing includes two stages as follows:
1) Rows with outlier data were deleted from the total data; in this regard, outlier is detected both in row and column forms. DFBETA algorithm was used in the row method to detect outliers (16).
2) In column deletion method (Equation 1) was applied for each column.matrix, which is calculated using Equation (4) (Figure2). It should be mentioned that *var* means variance. If the obtained quantities were less than minimum or higher than maximum, they would be regarded as outliers. Each line containing these outliers was deleted. In the second stage of preprocessing, data change within the range of 0-1 was evaluated (4). This process was repeated for all the columns.

### *1.6. Data deletion*

In this process, the rows were first divided into two parts of observation and missingness. The maximum missingness in rows was 50%, which was indicative of at least 50% data completion. Accuracy in the estimation of missing data depends on the type and amount of missing data. The amount of available missing data in a database is different; accordingly, simulation of the amount of missing data was conducted using various rates. Usually, the missing data rate is 1%-20% for producing missing data in a database. According to what was stated about 1% rate, missing data rates for 3%, 5%, 10%, 15%, and 20% can be achieved. However, the rate of missing data is just one of the important aspects in creating a database for missing data. The number of missing fields is determined with regard to the desired rate, which should be randomly assigned to the database records. Missingness is demonstrated based on the last records.

$$min = mean - \sqrt{var}\ ,\ \ max = mean + \sqrt{var} \quad (1)$$

$$Corr(X_i, X_j) = \frac{Cov(X_i, X_j)}{\sigma_i \sigma_j} = \frac{E[(X_i - \mu_i)(X_j - \mu_j)]}{\sigma_i \sigma_j} \quad (2)$$

$$Corr = \begin{bmatrix} Corr(X_1, X_1) & Corr(X_1, X_2) & \dots & Corr(X_1, X_m) \\ Corr(X_2, X_1) & Corr(X_2, X_2) & \cdots & Corr(X_2, X_m) \\ \vdots & \vdots & \ddots & \vdots \\ Corr(X_m, X_1) & Corr(X_m, X_2) & \cdots & Corr(X_m, X_m) \end{bmatrix} \quad (3)$$

$$Corr_{Distance} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} \left| Corr_{ij_{\text{observation}}} - Corr_{ij_{\text{observation+missing}}} \right|}{m^2} \quad (4)$$

$$\text{Minkowski Distance} = \left(\sum_{k=1}^{n} |X_k - Y_k|^r\right)^{\frac{1}{r}} \quad (5)$$

$$\text{Minkowski Distance} = \left(\sum_{k=1}^{n} |X_k - Y_k|^r\right)^{\frac{1}{r}} \quad (6)$$

$$\text{Similarity}_{i,1} = \frac{1}{L} \sum_{j=1}^{L} \frac{1}{\left(\sum_{k=1}^{n} \left| X_{i_k} - Y_{j_k} \right|^r\right)^{\frac{1}{r}}} \quad (7)$$

$$\text{Similarity}_{i,2} = \frac{1}{L} \sum_{j=1}^{L} \frac{1}{\left(\sum_{k=1}^{n} \left| X_{i_k} - Y_{j_k} \right|^r\right)^{\frac{1}{r}}} \quad (8)$$

$$\text{Similarity}_i = \left| \text{Similarity}_{i,1} - \text{Similarity}_{i,2} \right| \quad (9)$$

$$\text{Similarity} = \frac{1}{m} \sum_{i=1}^{m} \text{Similarity}_i \quad (10)$$

$$Fitness = \alpha Corr_{Distance} + \beta \text{Similarity} \quad (11)$$

$$\text{K\_ClusterNum} = \left\lfloor \frac{number}{\text{iteration} * 10} \right\rfloor \quad (12)$$

$$ELR = \alpha \times \frac{Number\ of\ current\ cucloo\ seggs}{Total\ number\ of\ eggs} \times (Var_{hi} - Var_{low}) \quad (13)$$

$$Index\ of\ agreement = \left(1 - \left[\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{\sum_{i=1}^{n} (|P_i - \bar{O}| + |O_i - \bar{O}|)^2}\right]\right) \times 100 \quad (14)$$

$$N = \frac{m \times k \times r}{100} \quad (15)$$

**Figure 2.** Equations related to the study

***1.7. Problem solving with the CS algorithm***

The missing data are predicted with the CS algorithm, afterward, its optimality was appraised by evaluation parameters. Stages of using the proposed Cuckoo algorithm are explained in the following parts of the article.

***1.7.1.*** *Determination of Cuckoo nest position*

The Cuckoo position is the missing data, which is within the range of 0-1. Cuckoo position is randomly assigned using normal distribution (17).

***1.7.2.*** *Production of the initial population*

Population size must be determined at this stage; that is, it should be demonstrated how many habitats must exist in a treasury. One of the most important factors for algorithm functionality is the population size. If the population size is too small, only a small part of search solution space would be evaluated, which could lead to immediate convergence to a local optimum. On the other hand, if the population is too big, a lot of calculation that is disproportionate to the obtained solution will be carried out, which eventually leads to prolonged implementation. The optimum initial population for longitudinal data of CD is five times bigger than the total number of missing data (this value is calculated by trial and error).

***1.7.3***. *Fitness function*

In the problem of solving missing data using a metaheuristic algorithm, determining the fitness function is one of the most important issues; the designed function entails two parts, as follows:

- Correlation coefficient is used in the first part, which estimated the relationship between two columns.
- The similarity feature is applied in the second part.

In the first stage, the rows of table are divided into two parts; the first part includes rows with no missing data, which is called observation data. In the second part, there are rows with at least one missing data, recognized as missing.

The missing values must be set in a way that no change is made to the relationship between data. In so doing, two

fitness function parameters of observation and missing data grow closer.

**1.7.3.1.** Correlation coefficient

Correlation coefficient is a statistical tool used to determine the type and degree of relationship between two quantitative variables. This tool is also a criterion applied to determine the correlation between two variables, which also demonstrates the strength and type (direct or reverse) of relationship. This factor is between 1 and -1, which is zero in case of lack of relationship between the two variables. The correlation between the two randomized variables of $X_i$ and $X_j$ is described using (Equation 2) (17). If the number of columns is considered as *m*, the correlation coefficient matrix of $m \times m$ is created as observed in (Equation 3). The correlation coefficient matrix is estimated twice, once for the observation data, which is called $Corr_{observation}$, and once for missing data using the predicted amounts. The correlation coefficients of total observation and missing data are calculated and called $Corr_{observation+missing}$. Thereafter, the less the difference between these two correlation coefficients, the more suitable the predicted values. This difference was calculated using Equation (4).

**1.7.3.2.** Similarity

Because similarity has a negative relationship with distance, the latter is estimated first. Afterward, similarity is considered equal to the reverse value of distance. For the two variables of *X* and *Y*, distance is calculated using the Minkowski equation, which is observed in (Equation 5). In (Equation 5), *r* is equal to 2 and similarity is estimated using (Equation 6). The following estimations are conducted to calculate the similarity between all rows of missing table. The similarity is estimated in two modes for the *i*th rows. In the first mode, all the columns are considered, and the similarity between the *i*'th row and other column rows is calculated. Afterward, the mean of the obtained value is regarded as the similarity of the first mode, estimations of which are provided in (Equation 7). The observation data are called *Y* in (Equation 7), and their number is equal to L. On the other hand, the missing data are considered as *X*, and the number of columns is recognized as *n*. The following estimations are conducted to calculate the similarity between all rows of missing table. However, the rows containing missing data are deleted in the second mode, followed by estimation of similarity. (Equation 8) is applied to estimate similarity; meanwhile, the only difference is reducing data due to elimination of rows with missing data of $\acute{n}$. Thereafter, the absolute difference value between $Similarity_{i,1}$ and $Similarity_{i,2}$ is calculated using Equation (9). Similarity is estimated for all rows of the missingness table, followed by calculation of mean similarities using Equation (10). In this equation, *m* is equal to the number of rows in the missing table.

***1.7.4.*** *Calculation of the fitness function*

To calculate the fitness function, a linear combination of the correlation coefficient and similarity is employed, as shown in (Equation 11). In the above-mentioned equation, the variables of α and β are demonstrated as α=1 and β=3 (these values are calculated by trial and error).

***1.7.5.*** *Determination of the algorithm parameters*

**1.7.5.1.** Lower and upper limits of the algorithm

With regard to the type of selected habitat, the minimum and maximum values of the algorithm are zero and 1, respectively.

**1.7.5.2.** The minimum and maximum laying

This parameter is one of the most important parts of the algorithm; the minimum and maximum values of which are calculated at 2 and 3 using trial and error.

**1.7.5.3.** Maximum number of generations

This parameter is estimated at 10 using trial and error.

**1.7.5.4.** Number of K-means clusters

In terms of structure of evolutionary algorithm, the algorithm must first be explored and then exploited. Therefore, the following equation was applied, in which *number* demonstrates the initial population of the algorithm (8).

**1.7.5.5.** Motor constant

Herein, the motor constant is first estimated at 1, from which 0.1 is subtracted in each generation and its amount reaches zero at the end.

**1.7.5.6.** Determining the optimal algorithm

The criteria for ending the algorithm is one of the following:

- Production of 100 generations without changing the fitness function, calculated with trial and error.
- At the end of the 100 generations

**1.7.5.7.** Determining the maximum Cuckoos for the next generation

Excessive definition of this parameters can slow down the algorithm, whereas its low definition leads to inadequate optimal response from the algorithm. Thus, it is considered 30% more than the initial population, which also was calculated via trial and error.

**1.7.5.8.** Laying radius coefficient
In nature, each cuckoo lays 5 to 20 eggs. This number is used as the upper and lower limits of allocating eggs to each cuckoo at different repetitions. Another habit of real cuckoos is laying eggs in a unique domain, known as the maximum laying domain (ELR8). In an optimization problem, each variable has an upper (varhi) and lower (varlow) limit, and each ELR could be defined using these limits. ELR is equal to the total number of eggs, including the current number of cuckoo eggs, as well as the upper and lower limits of the problem variables. Therefore, ELR is estimated using Equation (21). The maximum value of ELR is adjusted with alpha (18).

**2. Evaluation**
One of the most important parts of this model is conducting tests and confirming the obtained results. We created programs with MATLAB language to evaluate the proposed model, which will be further explained in the study. A fast-processing computer with the following characteristics was employed. It is worth mentioning that a fast-processing computer was selected due to the high volume of computations. Data on CD were used to evaluate the algorithm, containing 631 records and seven features (five quantitative and two categorical features), which are shown in Table 1. Evaluation of the data was conducted through the implementation of each program for 100 times due to the use of a probability-based algorithm. Afterward, outliers were deleted, and the mean of data was considered as the output.

**Table 1.** Summary of quantitative data in cervical dystonia database

| Statistical parameters | $Feature_1$ (Sex) | $Feature_2$ (Week) | $Feature_3$ (Site) | $Feature_4$ (Id) | $Feature_5$ (Age) | $Feature_6$ (Twstrs) | $Feature_7$ (Treat) |
|---|---|---|---|---|---|---|---|
| Minimum | 0.00 | 0.00 | 1.00 | 1.00 | 26.00 | 6.00 | 0.00 |
| First Quartile | 0.00 | 2.00 | 3.00 | 3.00 | 46.00 | 32.50 | 0.00 |
| Median | 0.37 | 4.00 | 6.00 | 7.00 | 56.00 | 43.00 | 0.99 |
| Mean | 0.00 | 6.957 | 5.119 | 7.021 | 55.62 | 41.48 | 1.00 |
| Third Quartile | 1.00 | 12.00 | 8.00 | 10.00 | 65.00 | 51.00 | 2.00 |
| Maximum | 1.00 | 16.00 | 9.00 | 19.00 | 3.98 | 71.00 | 2.00 |

***2.1. Evaluation criteria***
Different criteria were used to confirm the accuracy of results of the amount attributed to the missing value. A well-known evaluation criterion, identified as index of agreement, is applied to evaluate and compare results of the proposed method. In this regard, Equation (14) is used (19). In Equation 14, $N$ is the total number of missing members in data, which depends on the missing data rate and the size of database. A database with M rows, K columns, and r% missing rate is estimated using Equation (15). In Equation (14), the quantity of $O_i$ is equal to the actual quantity of ith member of the database. Mean of this quantity is represented with $\bar{O}$. In this formula, the assigned quantity is provided using the proposed method for each $O_i$ and is shown as $Pi$.

***2.2. Comparison of the results of the proposed method***
In this study, the proposed algorithm was used for data on the CD disease. In so doing, the missingness percentage is considered 1%–20%, and the missingness of the number of rows was 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, and 50%. In addition, evaluation of the algorithm is provided in Table 2. The percentages shown in the gray areas are indicative of the possibility of lack of missingness in the rows. As observed in Table 2, increased missingness in rows is accompanied with lower accuracy of the compatibility index. Moreover, higher percentage of missingness is associated with predicted accuracy of the compatibility index. In this regard, the maximum and minimum compatibility indices are 98.48% and 91.56179%, respectively, with the mean compatibility index of 94.90482±2.3402%.

**Table 2.** Evaluation of the compatibility index in the proposed algorithm for data on the cervical dystonia

| Total Missingness | Row Missingness | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 5% | 10% | 15% | 20% | 25% | 30% | 35% | 40% | 45% | 50% |
| 1% | 98.48 | 98.39 | 97.54 | 97.49 | 96.79 | 96.79 | 96.61 | 96.38 | 95.5 | 95.11 |
| 2% | - | 97.65 | 97.24 | 97.19 | 96.59 | 96.52 | 96.51 | 96.33 | 95.49 | 95.1 |
| 3% | - | 97.16 | 96.82 | 96.81 | 96.38 | 96.34 | 96.26 | 96.21 | 95.39 | 95.1 |
| 4% | - | - | 96.33 | 96.23 | 96.13 | 96.12 | 96.05 | 95.93 | 95.32 | 95.04 |
| 5% | - | - | 95.95 | 95.91 | 95.9 | 95.87 | 95.86 | 95.8 | 95.24 | 95.02 |
| 6% | - | - | - | 95.72 | 95.67 | 95.65 | 95.65 | 95.63 | 95.63 | 95.02 |
| 7% | - | - | - | 95.47 | 95.45 | 95.21 | 95.15 | 95.09 | 95.07 | 94.95 |
| 8% | - | - | - | - | 94.91 | 94.89 | 94.87 | 94.83 | 94.83 | 94.66 |
| 9% | - | - | - | - | 94.68 | 94.67 | 94.65 | 94.56 | 94.37 | 94.06 |
| 10% | - | - | - | - | - | 94.44 | 94.41 | 94.38 | 94.3 | 94.03 |
| 11% | - | - | - | - | - | 94.28 | 94.28 | 94.25 | 94.18 | 94.03 |
| 12% | - | - | - | - | - | - | 93.99 | 93.96 | 93.72 | 93.63 |
| 13% | - | - | - | - | - | - | 93.81 | 93.75 | 93.66 | 93.69 |
| 14% | - | - | - | - | - | - | - | 93.36 | 93.09 | 93.02 |
| 15% | - | - | - | - | - | - | - | 93.22 | 93.06 | 93.01 |
| 16% | - | - | - | - | - | - | - | 93 | 92.99 | 92.8 |
| 17% | - | - | - | - | - | - | - | 92.61 | 92.46 | 92.23 |
| 18% | - | - | - | - | - | - | - | 92.41 | 92.4 | 92.21 |
| 19% | - | - | - | - | - | - | - | 92.36 | 92.34 | 92.18 |
| 20% | - | - | - | - | - | - | - | 92.02 | 91.98 | 91.56 |

**3. Discussion and conclusions**

Currently, missingness is one of the most important issues of longitudinal studies and clinical trials. In this regard, the concepts and artificial neural algorithms were functionally evaluated in the present study. Missingness and its associated outcomes, patterns and mechanisms of this concept, negligible nature and suitable proposed models, and eventually proper models proposed for the un-negligible missing data were discussed in this study. During the recent years, significant software advances were made in the modeling of longitudinal data with missingness. Most common statistical software, such as SAS, S-plus, R, SPSS, and Stata, can run models with missingness. Because the missingness data always remain unobserved, all the present methods for evaluating these data include undetectable and unprovable assumptions. Therefore, when faced with missing data, we recommend conducting proper sensitivity evaluations and not relying on the results of just one method. Other research showed that meta-heuristic algorithms can properly impute lost data, generating suitable output with better performance (20, 21). In their research (22), it was shown that meta-heuristic algorithms have suitable output with adequate model accuracy. In a paper by Jiang et al. in 2016, the “Cuckoo search” algorithm was shown to be a proper method for placing the lost data (22). In this paper also, the Cuckoo search algorithm was utilized, and the optimality of its utilization for placing the lost data was proven. In a paper by Garren, it was indicated that the correlation coefficient is a proper method for placing the lost data (23). In this research, the correlation coefficient was used for placing the lost data in the “proposed algorithm” section. In the indicated model evaluation, the model was found to be suitable. In a paper by Junnenen et al., the “index of agreement” parameter was shown to be a suitable method for evaluating the lost data, which also was used in this research (19). In summary, it could be concluded that application of two methods of similarity and correlation coefficient can provide more proper results in predicting the missing values.

**Acknowledgments:**

This work was a part of the first author’s PhD dissertation in supported by a [grant #931034] from Mashhad University of Medical Sciences Research Council. The funder was involved in preparation and publication process of manuscript.

**Conflict of Interest:**

There is no conflict of interest to be declared.

**Authors' contributions:**

All authors contributed to this project and article equally. All authors read and approved the final manuscript.

**References:**


1) Wikipedia, the free encyclopedia. Longitudinal study. 2016. Available from: https://en.wikipedia.org/w/index.php?title=Longitudinal_study&oldid=731082311.
2) Enders CK. Applied Missing Data Analysis (Methodology in the Social Sciences). ISBN-13: 978-1606236390. Available from: https://www.amazon.com/Applied-Missing-Analysis-Methodology-ciences/dp/1606236393/ref=sr_1_cc_1?s=aps&ie=UTF8&qid=1471519943&sr=1-1-catcorr&keywords=Applied+Missing+Data+Analysis.
3) Mayo Clinic. Spasmodic torticollis. 2016. Available from: http://www.mayoclinic.org/diseases-conditions/spasmodic-torticollis/basics/definition/con-20028215.
4) 2016. Available from: http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets /Ccdystonia.html.
5) Nakai M, Chen DG, Nishimura K, Miyamoto Y. Comparative study of four methods in missing value imputations under missing completely at random mechanism. Open J Stat. 2014; 4: 27-37. doi: 10.4236/ojs.2014.41004.
6) Little A, Rubin B. Statistical Analysis with Missing Data. 3th ed. Chichester: John Wiley & Sons; 2016.
7) Yang XS, Deb S. Engineering optimisation by cuckoo search. Int J Math Model Numer Optim. 2010; 1(4): 330–43.
8) Payne RB. The Cuckoos. Oxford University Press; 1833.
9) Brown C, Liebovitch LS, Glendon R. Lévy flights in Dobe Ju/'hoansi foraging patterns. Hum Ecol. 2007; 35(1): 129–38. doi: 10.1007/s10745-006-9083-4.
10) Reynolds AM, Frye MA. Free-flight odor tracking in Drosophila is consistent with an optimal intermittent scale-free search. PloS One. 2007; 2(4): e354. doi: 10.1371/journal.pone.0000354. PMID: 17406678, PMCID: PMC1831497.
11) Pavlyukevich I. Lévy flights, non-local search and simulated annealing. J Comput Phys. 2007; 226(2): 1830-1844. doi: 10.1016/j.jcp.2007.06.008.
12) Shlesinger MF. Mathematical physics: Search research. Nature. 2006; 443(7109): 281–2. doi: 10.1038/443281a. PMID: 16988697.
13) Barthelemy P, Bertolotti J, Wiersma DS. A Lévy flight for light. Nature. 2008; 453(7194): 495-8. doi: 10.1038/nature06948. PMID: 18497819.
14) Reynolds AM, Frye MA. Free-flight odor tracking in Drosophila is consistent with an optimal intermittent scale-free search. PloS One. 2007; 2(4): e354. doi: 10.1371/journal.pone.0000354. PMID: 17406678, PMCID: PMC1831497.
15) Yang XS. Biology-derived algorithms in engineering optimization. In: Handbook of Bioinspired Algorithms and Applications. 2006.
16) Ford GS. Outlier Statistics Using Eviews. 2008. doi: 10.2139/ssrn.1293945.
17) Wheelan C. Naked Statistics: Stripping the Dread from the Data. 1 edition. W. W. Norton & Company; 2014.
18) Witten IH, Frank E, Hall MA. Data Mining: Practical Machine Learning Tools and Techniques. Third Edition. Burlington, MA: Morgan Kaufmann; 2011.
19) Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M. Methods for imputation of missing values in air quality data sets. Atmos Environ. 2004; 38(18): 2895-907. doi: 10.1016/j.atmosenv.2004.02.026.
20) Lobato F, Sales C, Araujo I, Tadaiesky V, Dias L, Ramos L, et al. Multi-objective genetic algorithm for missing data imputation. Pattern Recognit Lett. 2015; 68: 126-31. doi: 10.1016/j.patrec.2015.08.023.
21) Leke C, Marwala T, Paul S. Proposition of a Theoretical Model for Missing Data Imputation using Deep Learning and Evolutionary Algorithms. ArXiv Prepr ArXiv151201362. 2015.
22) Jiang P, Liu F, Wang J, Song Y. Cuckoo search-designated fractal interpolation functions with winner combination for estimating missing values in time series. Appl Math Model. 2016; 40(23): 9692-718. doi: 10.1016/j.apm.2016.05.030.
23) Garren ST. Maximum likelihood estimation of the correlation coefficient in a bivariate normal model with missing data. Stat Probab Lett. 1998; 38(3): 281–8. doi: 10.1016/S0167-7152(98)00035-2.