SIMILARITY ASSESSMENT FOR THE POPULATION DENSITY OF SOME CITIES IN NIGERIA USING DYNAMIC TIME WARPING ALGORITHMS

***Adedoyin, E.D[1], Awosusi, B.M[2], Ayodele, O.U[3] and Unuigbokhai, A.O[4]**
[1,2,3,4]Environmental Modeling and Biometrics Department
Forestry Research Institute of Nigeria
Forest Hill, Jericho, Ibadan
***Corresponding e-mail:** daveadedoyin2021@gmail.com; **Tel: +2347068420674**

ABSTRACT

Similarity, or proximity, measures are used in diverse fields of inquiry to study the trajectory of random variables. Patterns, or trajectories, of human population density growth (measured with respect to constant area) are well-documented to be closely associated with social, economic and environmental development and vulnerabilities. This study is therefore aimed at investigating the trajectory of the population density of some Nigerian cities having population density ≥ 900 persons.km^{-2} with a view of clustering the cities using Dynamic Time Warping (DTW) algorithms as the distance measure. The cities considered were selected on the basis of 2006 National Population Commission Census' report. A Preliminary investigation for the optimal cluster number using K-means, Partition Around Medoids (pam) and Agglomerative Hierarchical algorithms showed that k = 2 and k = 6 produced optimal clusters. Since a higher optimal cluster value connote production of better grouping, outputs for k = 6 was selected. The result showed a dissimilar population density trajectory for Okene and Zaria while Uyo and Ikorodu cities had similar population density trajectory for the periods considered. Although Kano showed similar population density trajectory with Aba and Enugu, the cities of Enugu and Aba had more similar density trajectory than the city of Kano.

Keywords: Population Density, Clustering, Similarity, Time Series

INTRODUCTION

Population density is one of the most important indices of urbanization. It holds vital information on the well-being, growth, socioeconomic development and environmental dynamics of a geographical space, particularly when combined with some important indices in these areas consequently aiding good planning and monitoring. For instance, Hummel (2020) investigated the "effects of population and housing density in urban areas on income in the United States utilizing the relationship between urban density and incomes". In this study, population density or urban density was found to result in higher social interaction and knowledge spill-over leading to skills and opportunities exchange and consequently resulting to increased income. Hazarie *et al*. (2021) affirmed population density to lead to increased social interaction but noted that an increased social interaction often resulted to faster infectious disease transmission and epidemic outbreaks. de Sherbinin *et al.* (2007) on the other hand related population density to environmental transformation and climate change.

Since urbanization and population density are intertwined, an analysis of the characteristics of population density (e.g. in terms of its trajectory) could reveal important information on a city's growth pattern. It could also assist in the design of appropriate policies and as well guide planning strategies and, in some cases, form a good platform for adapting an existing policy and/or strategy where two or more cities present similar density trajectory.

Population density observations, when collected over time, are time series. Time series data however present a number of challenges (e.g. high-dimensionality, autocorrelation dynamism etc.) and are widely occurring in different fields of applied Statistics. Some of these fields are such that perform clustering (e.g. in Remote Sensing and GIS for image time series clustering to aid land use and cover monitoring (Maus *et al.*, 2016)), classification (e.g. in bioinformatics and medical studies in gene expression data matching (Yuan *et al.,* 2011)), similarity search (e.g. in economics and finance for identifying business cycles similarities (Franses and Wiemann, 2020)) or forecasting of time series observations.

Time series clustering is an unsupervised data mining technique for organizing data points (or observations) into groups (called clusters) based on some similarity measure. These aggregated observations are such that the similarity of objects in the same cluster is maximized while it is minimized between objects existing in different clusters (Iglesias and Kastner, 2013). Studies on similarity of objects have shown that similarity measurements are distance-measure dependent. Some of the most commonly used distance functions are Manhattan distance function, Euclidean distance function, Chebyshev distance function and Pearson's correlation coefficient, amongst others. Investigation on the performance of these measures on varied dataset have shown that these distance measures may not perform optimally with time series. For instance, though the Pearson's correlation coefficient may be used for finding similarity of variables, it is sensitive to the presence of outliers and fails when two time series variables are out of phase, often exhibiting similarity when there is dissimilarity. The Euclidean distance, like the Pearson's correlation coefficient, cannot handle two series which are out of phase (Novak and Mirshahi, 2021). These problems are however well handled by Dynamic time warping (DTW) distance algorithm. DTW algorithm is a nonparametric tool designed for assessing similarity between two time series variables. This study is therefore aimed at identifying the similarity of population density trajectory of some Nigerian cities with DTW algorithm used as the distance function. This is of great importance to researchers, development managers and policy makers. For instance, the outcome of this study could be used as a prerequisite for further in-depth comparative studies (e.g. exposure to some risk factors within similar cities etc.). It may also aid developmental designs for allocation of transportation, services and environmental facilities (e Silva *et al.*, 2020) as well as guide policy formulation.

METHODOLOGY

Data Description

Data on population density used in this study was obtained from macrotrends website, a site which sources its dataset from the World Population Prospects of the United Nations. The data comprise annual population density for sixteen Nigerian cities, mostly long-existing (Aliyu and Amadu (2017), from 1950 to 2020. Each city selected was one of the seventeen cities with the highest population density (i.e. population density $\geq 900\ people.km^{-2}$) in Nigeria as identified by the National Population Commission's (NPC) 2006 Census report. Though Maiduguri was listed one of the seventeen cities, it was excluded because of the on-going insurgency attacks in Borno State– the state which houses the city of Maiduguri. The cities considered were: Aba, Lagos, Kano, Kaduna, Onitsha, Oshogbo, Port Harcourt, Katsina, Ikorodu, Enugu, Zaria, Ikot Ekpene, Uyo, Ado Ekiti, Ilorin and Okene.

Dynamic Time Warping (DTW) Algorithm

DTW algorithm is used for computing the distance and alignment between two time series (Seto *et al.,* 2015). The process for this algorithm follows three basic steps:

STEP 1 (Derivation of Alignment for Points)

Let $X_i = \{x_{i1}, \dots, x_{is}, \dots, x_{im}\}$ and $X_j = \{x_{j1}, \dots, x_{jt}, \dots, x_{jn}\}$ be two time series of length m and n, respectively, such that X_i and $X_j =$ population density in city i and city j, for $i \neq j$, at time points s and t ($s \in m, \ t \in n$). Also, let a local cost distance matrix, denoted by C, of dimension $m \times n$ be defined on the points of X_i and X_j such that the elements of C are the cost function, c, which is defined as:

$$C(i,j) = c(x_i, x_j) = \left(\sum \left|x_i - x_j\right|^p\right)^{\frac{1}{p}}. \qquad (1)$$

Each matrix element (i,j) in Equation (1) then corresponds to the alignment between the points x_i and x_j. Entries in parenthesis on RHS of Equation (1) is the generalized Minkowski distance bearing a value of $p = 1$ when Manhattan distance is to be used and $p = 2$ for Euclidean distance.

STEP 2 (Derivation of the warping path)

The DTW algorithm then finds the warping path that minimizes the alignment between X_i and X_j by iteratively stepping through the local cost distance matrix, $C(i,j)$. A warping path

$$W = (w_1, \ w_2, \dots, w_K), \qquad \max(m,n) \leq K < m + n - 1,$$

is a contiguous set of matrix elements defining a mapping between X_i and X_j such that the kth element of W is defined as $w_k = (i,j)_k$ and the path, W, satisfies:

i. Boundary condition: $w_1 = (1,1)$ and $w_K = (m,n)$;

ii. Continuity condition: Given $w_k = (a,b)$, then $w_{k-1} = (a', b')$, where $a - a' \leq 1$ and $b - b' \leq 1$.

iii. Monotonicity condition: Given $w_k = (a,b)$, then $w_{k-1} = (a', b')$, where $a - a' \geq 0$ and $b - b' \geq 0$.

STEP 3 (DTW Measure)

The DTW distance is then the sum of the pointwise distances along the optimal path, W*, for the cost function $c(w)$ such that (Seto e*t al.*, 2015):

$$DTW(X_i, X_j) = \min_W \left\{\sum_{k=1}^{K} w_k\right\}. \qquad (2)$$

The optimal alignment is then given by the warping path, p, that minimizes the cumulative distance of all mapped point pairs, computed recursively, by the recurrence relation:

$$\gamma(i, \ j) = d(x_i, x_j) + min\{\gamma(i-1, j-1), \gamma(i-1, \ j), \gamma(i, \ j-1)\} \qquad (3)$$

where $d(x_i, x_j)$ is the distance measured between points x_i and x_j and the ith start condition is $d(0, \ 0) = 0$.

Hopkins Clustering Tendency Statistic (H)

Hopkin’s statistic (Brian and Skellam, 1954) is used the clustering tendency of a dataset. In other words, this measure investigates whether a given dataset contains meaningful clusters. The Hopkin’s statistics is defined as:

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}. \tag{4}$$

H = Hopkin’s statistic;

x_i = distance between each point in the dataset and a corresponding nearest neighbour;

y_i = distance between each point in the simulated dataset with same variance as the original dataset and a corresponding nearest neighbour;

n = sample size.

The null hypothesis for this statistic (H) is given as:

$$H_o{:}\, The\ dataset\ D\ is\ uniformely\ distributed\ (i.e.no\ meanful\ clusters).$$

This hypothesis is rejected when the values of H is close to zero.

Hierarchical Cluster Algorithm

The hierarchical agglomerative clustering algorithm was used for partitioning the cities into different clusters. This task was done iteratively using the dynamic time warping distance measure and a complete linkage method. The complete linkage method aggregates objects such that the distance between two clusters is the maximum distance between their objects.

Cluster Evaluation

The population density clusters were evaluated using seven internal validation indices (Table 1). These indices measure the goodness (i.e. quality as well as usefulness) of the generated clusters using information from the clusters and their objects; that is, no external information were used. All analysis were done within the R programming environment, version 4.1.2.

Name	Notation	Formula
Silhouette index (Rousseeuw, 1987)	Sil	$\frac{1}{NC}\sum_i \frac{1}{n_i}\sum_{x\in C_i} \frac{\min_{j\neq i}\left[\frac{1}{n_j}\sum_{y\in C_j} d(x,y)\right] - \frac{1}{n_i-1}\sum_{y\in C_i, y\neq x} d(x,y)}{\max\left[\left(\min_{j\neq i}\left[\frac{1}{n_j}\sum_{y\in C_j} d(x,y)\right]\right),\ \left(\frac{1}{n_i-1}\sum_{y\in C_i, y\neq x} d(x,y)\right)\right]}$
Score function index (Saitta *et al.*, 2007)	SF	$1 - \frac{1}{e^{e^{\left\{\left(\frac{\sum C_k\in C \lvert C_k\rvert d(\overline{c_k},\bar{X})}{N\times K}\right)+\left(\sum C_k\in C \frac{1}{\lvert C_k\rvert} \sum x_i\in C_k\, d(x_i,\ c_k)\right)\right\}}}}$
Calinski-Harabasz index (Calinkski and Harabasz, 1974)	CH	$\frac{\sum_i n_i d^2(c_i,c)/(NC-1)}{\sum_i \sum_{x\in C_i} d^2(x,c_i)/(n-NC)}$
Davies-Bouldin index (Davies and Bouldin, 1979)	DB	$\frac{1}{NC}\sum_i \max_{j\neq i}\left(\frac{\frac{1}{n_i}\sum_{x\in C_i} d(x,c_i) + \frac{1}{n_j}\sum_{x\in C_j} d(x,c_j)}{d(c_i,c_j)}\right)$
Modified Davies-Bouldin index (Kim and Ramakrishna, 2005)	DB*	$\frac{1}{K}\sum_{C_k\in C}\frac{\max_{c_l\in C\backslash c_k}\{S(c_k)+S(c_l)\}}{\min_{c_l\in C\backslash c_k}\{d(\overline{c_k},\bar{c}_l)\}}$
Dunn index (Dunn, 1974)	D	$\min_i\left\{\min_j\left[\frac{\min_{x\in C_i, y\in C_j} d(x,y)}{\max_k\left\{\max_{x,y\in C_k} d(x,y)\right\}}\right]\right\}$

COP index (Gurrutxaga *et al.*, 2010)	COP	$\frac{1}{N} \sum_{c_k \in C} \lvert c_k \rvert \frac{\frac{1}{\lvert c_k \rvert} \sum_{x_i \in c_k} d(x_i,\overline{c_k})}{\min_{x_i \notin c_k} \max_{x_j \in c_k} d(x_i,x_j)}$

Table 1: Internal cluster validation measures used (**n:** number of objects in dataset (D); **c:** center of D; **NC:** Number of clusters; **C$_i$:** the ith cluster; **n$_i$:** number of objects in C$_i$; **c$_i$:** center of C$_i$; **d(x, y)**: distance between x and y)

RESULTS AND DISCUSSION

An exploratory analysis of the trajectory of each city's population density (Figure 1) showed that each density was growing exponentially with Onitsha showing a clearly distinct growth rate from 2000 to 2020. Analysis of the clustering tendency of the dataset using Hopkin's test statistic (H) gave a 0.2103 value which was significantly away from the 0.50 limit and closer to a zero value. Therefore, the null hypothesis of the test was rejected thereby upholding the alternative hypothesis that the dataset considered in this study can be aggregated into meaningful set of clusters.

Assessment of the optimal number of clusters (denoted k) that could be used for aggregating the objects using the "NbClust" function in R showed that optimality was attained at k = 2 and k = 6. This number of clusters were obtained to be agreed to by the hierarchical clustering method, partition around the medoids method and kmeans algorithms. Since k = 2 and k = 6 both agreed to produce optimal results, k = 6 was used for data clustering using the hierarchical agglomerative clustering method and the dynamic time warping distance algorithm. The outcome of the clustered population density trajectory is presented on a Dendrogram (Figure 2).

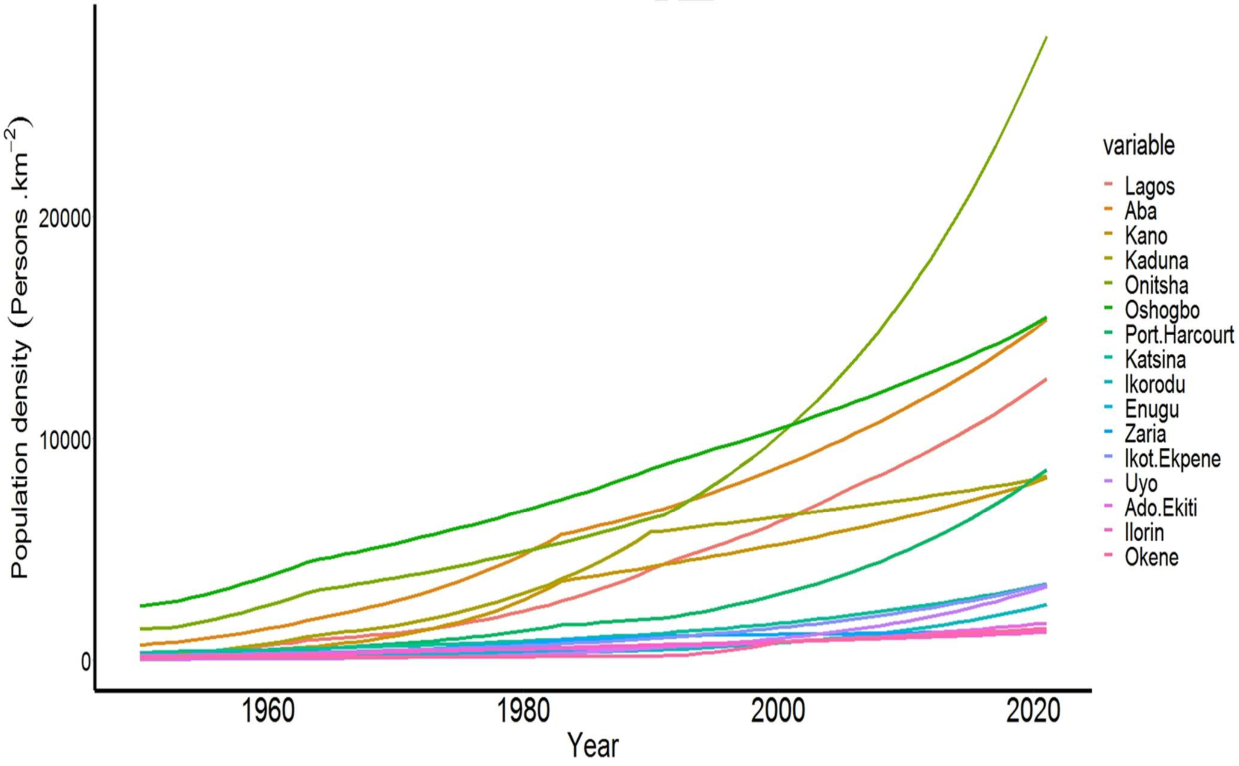

Figure 1: Population density trajectory of Nigeria's most densely cities (1950 – 2020)

Figure 2 showed the entire sixteen city population density trajectories partitioned into two major groups with each group further splitted to contain cities which share more similar population density trajectories. The Dendrogram showed that Zaria and Kaduna showed similar population density trajectory while the population density trajectory across the periods study was similar for Ilorin and Oshogbo cities. Ikorodu and Uyo were also observed to have similar population density trajectory. Although Onitsha, Port Harcourt and Ikot Ekpene showed similar population density trajectory, the study showed that Onitsha city and Port Harcout city had more similar population density trajectory than Ikot Ekpene city.

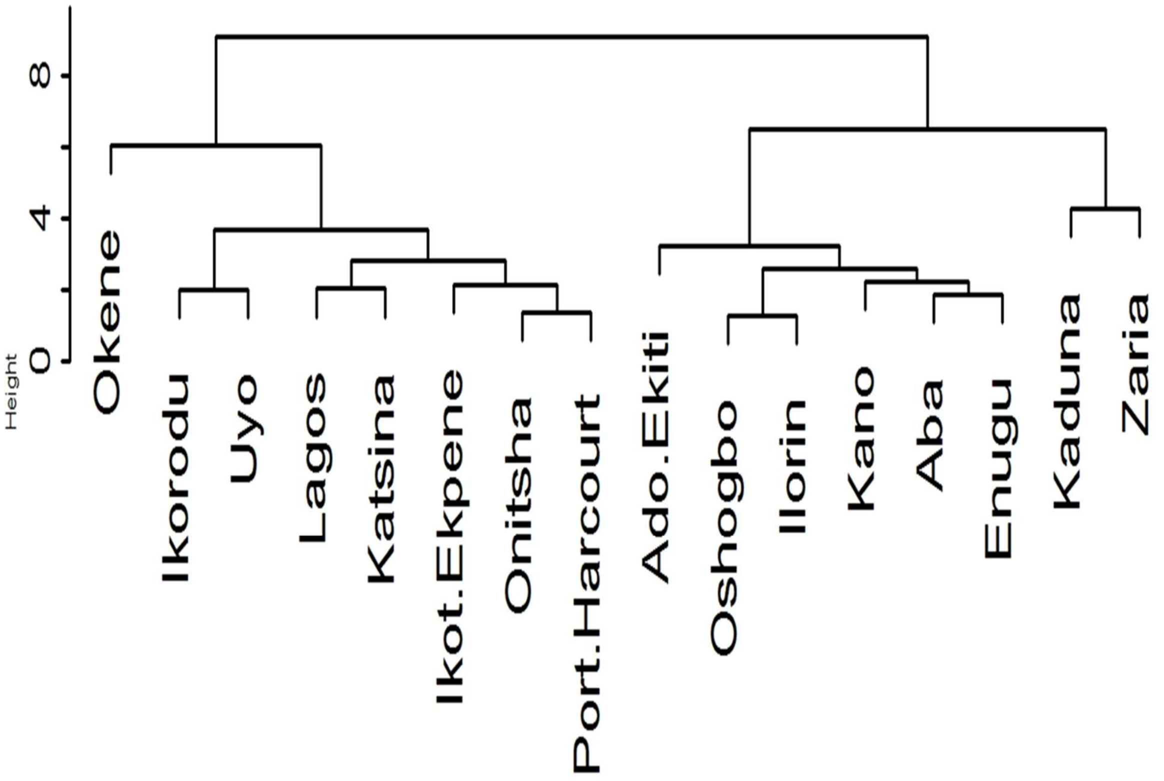

Figure 2: Dendrogram for trajectory of Nigeria city's population density

A table of the cluster validation index values for the seven evaluation measures used in this study is presented as Table 2. Values in bold face were suggested to give optimal cluster by the indices. Among the five clusters (k = 2, 3, 4, 5, and 6) tested against the population densities, the result showed k = 2, 5 and 6 to be preferred. This thus confirms the result presented by the NbClust function in the R function consequently implying that the use of k = 6 was sufficient and reliable.

CVI	K = 2	K = 3	K = 4	K = 5	K = 6
Sil	**0.52789**	0.45305	0.40460	0.31246	0.26871
SF	0.03106	0.00503	0.00598	**0.04857**	0.01921
CH	**20.27430**	12.16530	9.36867	8.03563	6.89850
DB	0.68234	1.01071	0.83030	**0.45962**	0.62333
DB*	0.68234	1.14694	0.90234	**0.68696**	0.81365
D	0.28127	0.46090	0.52393	0.52393	**0.56555**
COP	0.24580	0.18904	0.16125	0.14434	**0.12274**

Table 2: Cluster Validation Indices for Nigeria city's population density trajectory **(Values in boldface gives optimal cluster number by index)**

CONCLUSION

This study investigated the trajectory of the population density of the most densely populated cities in Nigeria, as identified by the NPC 2006 report. All city's population density trajectory for the study period were observed to increase exponentially while Onitsha had the most increase from Year 2000. Clusters with meaningful objects were derived from the data at k = 6. In the light of this study, the trajectory of the growth of the population density of some Nigeria cities can be grouped into identical groups using clustering method.

REFERENCES

Aliyu, A.A. and Amadu, L. (2017). Urbanization, cities and health: The challenges to Nigeria-A review. *Ann Afr Med*, 16(4): 149–158.

Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics,* 3(1): 1–27. Doi: 10.1080/03610927408827101.

Davies, D.L. and Bouldin, D.W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2): 224–227. Doi: 10.1109/tpami.1979.4766909.

Dunn, J.C. (1974). Well-separated clusters and optimal Fuzzy partitions. *Journal of Cybernetics*, 4(1): 95–104. Doi: 10.1080/01969727408546059.

Franses, P.H. and Wiemann, T. (2020). Intertemporal similarity of Economic time series: An) application of Dynamic Time Warping. *Computational Economics*, 56:59–75. https://doi.org/10.1007/s10614-020-09986-0.

Gurrutxaga, I., Albisua, I., Arbelaitz, O., Martín, J.I., Muguerza, J., Pérez, J.M. and Perona, I. (2010). SEP/COP: An efficient method to find the best partition in hierarchical clustering based on a new cluster validity index. *Pattern Recognition,* 43(10): 3364–3373. Doi: 10.1016/j.patcog.2010.04.021.

Hazarie, S., Sariano-Paños, D., Alex, A., Gómez-Gardeñes, J. and Ghosal, G. (2021). Interplay between population density and mobility in determining the spread of epidemics in cities. *Communications Physics,* 4:191. https://doi.org/10.1038/s42005-021-00679-0.

Hopkins, B. and Skellam, J.G. (1954). A new method for determining the type of distribution of plant individuals. *Annals of Botany,* XVIII (70): 213–228.

Hummel, D. (2020). The effects of population and housing density in urban areas on income in the United States. *Local Economy*, 35(1): 27–47. DOI: 101177/0269094220903265.

Iglesias, F. and Kastner, W. (2013). Analysis of similarity measures in time series clustering for the discovery of building energy patterns. *Energies,* 6: 579–597. doi: 10.3390/en6020579.

Kim, M. and Ramakrishna, R.S. (2005). New indices for cluster validity assessment. *Pattern Recognition Letters*, 26(15): 2353 – 2363. Doi: 10.1016/j.patrec.2005.04.007

Maus, V., Cĉmara, G., Cartaxo. R., Sanchez, A., Ramos, F.M. and de Queiroz, G.R. (2016). A time-weighted dynamic time warping method for land-use and land-cover mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* DOI: 10.1109/JSTARS.2016.2517118.

Novak, V. and Mirshahi, S. (2021). On the similarity and dependence of time series. *Mathematics,* 9, 550. https://doi.org/10.3390/math9050550.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65. Doi: 10.1016/0377–0427(87)90125–7.

Saitta, S., Raphael, B. and Smith, I.F.C. (2007). A bounded index for cluster validity. In: International Workshop on Machine Learning and Data Mining in Pattern Recognition, Heidelberg, Germany (pp. 174–187). Springer. Doi: 10.1007/978-3-540-73499-4_14.

Seto, S., Zhang, W. and Zhou, Y. (2015). Multivariate time series classification using Dynamic Time Warping template selection for Human activity recognition. arXiv:1512.06747v1.

de Sherbinin, A., Carr, D., Cassels, S. and Jiang, L. (2007). Population and environment. *Annu Rev Environ Resour.*, 32:345–373. DOI:10.1146/annurev.energy.32.041306.100243.

e Silva, F.B., Freire, S., Schiavina, M.,Rosina, K., Marin-Herrera, M.A., Ziemba, L., Craglia, M., Koomen, E. and Lavalle, C. (2020). Uncovering temporal changes in Europe's population density patterns using a data fusion approach. *Nature Communication*, 11:4631. https://doi.org/10.1038/341467-020-18344-5.

Yuan, Y., Chen, Y.P., Ni, S., Xu, A.G., Tang, L., Vingron, M., Somel, M. and Khaitovich, P. (2011). Development and application of a modified dynamic time warping (DTW-S) to analyses of primate brain expression time series. *Bioinformatics*, 12:347. http://www.biomedcentral.com/1471-2105/12/347.