On the Cluster Validity Test (s) in Unsupervised Machine Learning TDA Approach for Atmospheric River Patterns on Flood Detection in Nigeria

Felix Obi Ohanuba (felixohanuba@gmail.com)
Universiti Sains Malaysia https://orcid.org/0000-0003-3408-8568
Mohd Tahir Ismail
Universiti Sains Malaysia
Majid Khan Majahar Ali
Universiti Sains Malaysia
Ekele Alih
The Federal Polytechnic Idah
Precious Ndidiamaka Ezra
University of Nigeria Faculty of Physical Sciences

Research Article

Keywords: clustering, extreme climate, flood menace, machine learning, topology, big data, sustainable de-velopment goal (SDG)

Posted Date: June 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-459258/v1

On the Cluster Validity Test (s) in Unsupervised Machine Learning TDA Approach for Atmospheric River Patterns on Flood Detection in Nigeria

F. O. Ohanuba[a,b]*, M. T. Ismail[a], M. K. Majahar Ali[a] , E. Alih[c] and P.N. Ezra[b]

[a]*School of Mathematical Sciences, Universiti Sains Malaysia,11800 Penang, Malaysia;*

[b]*Faculty of Physical Sciences, Department of Statistics, University of Nigeria, Nsukka, Nigeria;*

[c]*Department of Mathematics and Statistics, Federal Polytechnic, Idah, Kogi State*

Correspondence details: Felix Obi, Mohd Tahir Ismail and, Majid Khan Majahar Ali

Felix O. Ohanuba is a PhD student, School of Mathematical Science, Universiti Sains Malaysia, (email: *felix.ohanuba@student.usm.my); Mohd T. Ismail is a Professor of Statistics, School of Mathematical Science, Universiti Sains Malaysia (email: m.tahir@usm.my), Majid K. Majahar Ali is a Senior Lecturer, School of Mathematical Science, Universiti Sains Malaysia (email: majidkhanmajaharali@usm.my); Ekele Alih is a Senior Lecturer, Department of Mathematics and Statistics, Federal Polytechnic, Idah, Kogi State (ekelson200@yahoo.com). Precious N. Ezra is a Lecturer, Faculty of Physical Sciences, Department of Statistics, University of Nigeria, Nsukka, Nigeria (email: precious.ezra@unn.edu.ng).

On the Cluster Validity Test (s) in Unsupervised Machine Learning TDA Approach for Atmospheric River Patterns on Flood Detection in Nigeria

Abstract

Abstract

TDA (i.e., Topological Data Analysis) has recently been a reliable and current research area in Statistics for extracting shape (information) from data. In this study, the researchers proposed an automated method that uses TDA & ML in identifying floods (ARs) in big data. Our process gives vital details on time series trends, which help mitigate the negative effect of ARs, such as flooding. The spatial data (between 1970 - 2018) from Nigeria Hydrological Services Agency (NIHSA) on four weather parameters were used. The daily datasets were converted to monthly datasets before the proposed method was applied. Python Software is used to develop code in the implementation of our process. Mostly, the outcome facts studied will drastically reduce disasters due to extreme events like floods and achieve some SDG goals related to the flood. The second objective is to identify potential flooding and no flooding in each zone. The work successfully used a real dataset and four variables that other studies have not used to fill a gap. After our model's training process, we obtained the best group at k = 2, where we have the highest Silhouette coefficient in each of the seven states. We have found a reasonable structure in the study considering the total average range (0.3 - 0.8). That gives an efficiency outcome of approximately 80%. Summary of clustered feature pattern shows the potential flood zone and no flood zone. We conducted cluster validity of our results using R software codes and, the test validated the best group at the same cluster k = 2. The Gap statistic shows efficiency ranging between 65% to 80% in the seven states. We found from **figure 11** that only the Silhouette plot obtained optimal values at exactly k = 2; The researchers got the extent of the spread from the centroid using Excel software.

*Keywords***:** clustering; extreme climate; flood menace; machine learning; topology; big data; sustainable development goal (SDG)

1. Introduction

The recognition of patterns of weather that usually leads to extreme events is challenging. It is the first step in understanding how they will vary in different climate change environments. Recognizing such events in climate and weather-prone communities is rarely studied. Extreme weather and climate change have brought about financial havoc, economic havoc, and human losses. Cities located in severe weather and climate zones are adversely affected almost every year. Many researchers have explored flood control, applying Topological Data Analysis (TDA) to study the pattern recognition using complex data (mostly simulated). Only a few researchers have recently involved TDA and Machine Learning (ML) in learning flood patterns. No study has used real data (data from the ground instrument) from reviewed works. Also, to the best of our knowledge, and from the reviewed literature, no study has used up to four weather parameters to study flooding. Our analysis also measured the extent of spread (variance) of the k=2 clusters.

TDA uses tools in mathematics that is concerned with studying shape (Wasserman 2018). A kind of geometry in which a foldable object (body) can be stretched or twisted can recover its original form when it relaxes. Topology originated in the 18th century by a Swizz Mathematician, Leonard Euler (Richeson 2019). (Wasserman 2018) stated that Topological Data Analysis (TDA) is "a collection of data analysis methods that finds structure in data. It includes clustering, manifold estimation, nonlinear dimension reduction, ridge estimation, and persistent homology" (Edelsbrunner 2013). TDA appeared gradually over time, and for historical development, see the following literature: (Frosini 1990, Robins 1999, Letscher and Zomorodian 2002, Carlsson et al. 2005, Zomorodian and Carlsson 2005). For further reading on the application of TDA and Machine Learning, see (Pascucci et al. 2010, Guiang and Levine 2012, Abou El Majd et al. 2018, Alaa and Mohamed 2017, Bubenik 2015, Carlsson et al. 2005). Persistent homology in the TDA method was applied to the human activity recognition task, which shows an 18.6% improved performance than the ordinary methods (Umeda, Kaneko, and Kikuchi 2019).

TDA refers to a large class of data analysis methods that uses notions of shape and connectivity. The advantage of taking this broader definition of TDA is that it will provide more context for recently developed methods (Wasserman 2018). Although clustering has the most straightforward version of TDA, we introduced a new TDA approach in solving the problem of floods. In this study, we used python software in writing a suitable code for the analysis. We also used codes in R software to evaluate the validity of our results. The validity test conducted shows the efficiency

of our method. In comparison, the Gap statistic is better than the Silhouette width in terms of connectedness and outcome result. Still, Silhouette seems better for indicating the exact number of clusters (i.e., k = 2) at the optimum plot. The idea of validating approach is "to standardize the graph by comparing it with its expectation under an appropriate null reference distribution of the data" (Tibshirani, Walther, and Hastie 2001). Plots of the features were conducted and, partitioning of the groups for the k number of clusters is also obtained.

(Guo and Banerjee 2017) highlighted the problem emanating from redundancy in sensor measurements and cited the use of multivariate statistical process control (MSPC) such as principal component analysis (PCA), Partial least-squares (PLS), and commented that "the two methods have served as the dominant ways of addressing the problem." (Krim, Gentimis, and Chintakunta 2016) introduces the common tools in applied algebraic topology, in an easily applicable context and offer a framework that naturally adapted to signal processing problems with some tools from linear algebra, examples and illustrations.

(Khasawneh, Munch, and Perea 2018, Riihimäki et al. 2020) "TDA" s classifier performed best in the other applications; the algorithm and software are beneficial as both a high combination of supervised machine learning technique and TDA alleviate the difficulties in choosing parameters with classification algorithms. TDA uses jointly considered features in the geometry and topology from complex data; it often preserves complex relationships. These properties made the technique unique and resulted in high performance in real-life applications. An important finding related to our expected finding and recommendation is that the TDA classifier achieves better accuracy when the data point (i.e., an increase of variable as we recommended) increases. Evidence is shown in (Riihimäki et al. 2020).

(de Gois et al. 2020) applied "normality and homogeneity tests in a 71-year time series of rainfall" (in Rio de Janeiro) and identified the most represented statistical tests in evaluating the collected dataset; the tests used only one variable, although a different test was conducted. (Chevyrev, Nanda, and Oberhauser 2018) TDA tools and Packages built-in R, such as Python sci-kit learn-package and vector classifier; the two-step operations result achieved advanced results on standard classification benchmarks. The approach Replicating Statistical Topology (RST) provides TDA on a more solid statistical background, which tackles one of the limitations of TDA in terms of the inability to make scientific claims that are statistically verifiable. (Gholizadeh and Zadrozny 2018) reviewed some areas of application of TDA in financial time series and system analysis; the paper

mentioned persistent homology (PH) as one of the techniques in TDA for finding patterns and forecasting. (Gidea and Katz 2018) the Combined TDA technique & the k-clustering approach and reviewed that the automatic classifier can identify time-series clusters (of a topologically distinct regime in crypto-currencies) for a specific period. PH, a morse-smale-based clustering algorithm, k-clustering, mapper clustering & hierarchical clustering with their R functions is robust when applied in measuring IQ in the gifted study; the results across various intelligence measurements have relative values compared with different TDA tools (Farrelly 2017). (Guiang and Levine 2012) pointed out that "classification is possible when the spectral data are sampled from the first wave-length range"; and that "one limitation with an application of TDA as a classifier (used in remote sensing) is the presence of contaminated pixels in the training set."

(Offroy and Duponchel 2016) explains how measuring shape and representation of the same are the two main tasks TDA performs; its important idea is that it considers a data set to be a point cloud (a sample) taken from a manifold in some high-dimensional space. (Offroy and Duponchel 2016) revealed the reason why topology is well suited for the analysis of big data sets in many areas; the three properties are a) coordinate invariance, b) deformation invariance and, c) compressed representation. Simplices are constructed from the sample data, and that simplices develop intervals, which joined together to form a kind of wireframe approximately of the manifold (Offroy and Duponchel 2016).

In the python code for data representation, we integrated multivariate statistical process control (i.e., the PCA and confusion matrix). For instance, the PCA serves as a primary means of addressing excessive measurements among sensors by "identifying and filtering out existing correlations in datasets that are erroneous, heterogeneous, and high-dimensional" (Guiang and Levine 2012). The confusion matrix in the code takes care of classifying results for the ML classifier (in percent) about testing accuracy during data preprocessing. We identify various flooding patterns in the seven selected regions in Nigeria and classify the features into flooding and non-flooding from the collected data to address the flooding problem. (Muszynski et al. 2019) stated that atmospheric rivers (ARs) are necessary for flood forecasts and that "ARs are long and narrow filaments of highly concentrated water vapor in the lower troposphere." ARs constitute more than 90% of water vapor transportation. It poses a high risk (extreme flooding, landfall, large rainfall, etc.) to society because of the large amount of data that a single AR can transport: AR occurs mostly in the world's coastal regions and causes enormous danger to the areas (Muszynski et al. 2019). (Echendu 2020)

focused on "the impact of flooding on Nigeria" and enumerated the specific SDGs most directly affected. Some SDGs, such as low or nonexistence drainage systems, flawed waste management systems, unregulated erecting of structures, and weak planning implementation, are mostly man-made. The work suggested a collaborative effort by concerned stakeholders to design and implement adequate Flood Risk Management (FRM) strategies. (Echendu 2020) suggested that the solution will include spatial planning and infrastructure planning.

According to (Echendu 2020), the SDGs related to flooding are as follows: (i) Sustainable development goal 1: "Zero poverty in all its forms everywhere, (ii) Sustainable development goal 2: "Terminate hunger, achieve food security and, improved nutrition and promote sustainable agriculture", (iii) Sustainable development goal 3: "Promote healthy living and well-being for all ages", (iv) Sustainable development goal 4: "Ensure inclusive and equitable standard education and promote durable learning opportunities for everyone", (v) Sustainable development goal 6: "Create availability and sustainably manage forest, combat dissertation, and also, to halt and reverse land degradation and avert biodiversity loss" (vi) Sustainable development goal 8: "Provide decent work and economic growth" (vii) Sustainable development goal 11: "To make cities and human settlements conducive, safe, and sustainable" (viii) Sustainable development goal 14: "To conserve and sustain the proper use of oceans, seas, and marine resources for sustainable-development (life on water)", (ix) Sustainable development goal 15: "To protect, restore and advocate sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and avert biodiversity loss (life on land)" (Echendu 2020)**.** It is necessary to mention that the first step in tackling the menace of flooding, including those mentioned in SDGs, is to recognize the flood pattern and see how the feature will react in the environment; this study achieved the objective in the succeeding results.

This study also fills a research gap using four variable parameters of real data and set to apply the TDA technique in the study of flood control in Nigeria. The aim of calculating the variance in **table 4** is to measure the spread from the mean value. The variance value is computed using Microsoft Excel to ensure minimal error in the calculation.

1.1 Problem formulation

"Nigeria is a country located in West Africa. It shares land borders with the Republic of Benin in west, Chad and Cameroon in the east, and Niger in the north" (Ngo 2019). "It also shares a border

in the southeast with the state of Ambazonia. Its coast lies on the Gulf of Guinea in the south, and it borders Lake Chad to the northeast" (Ngo 2019).

Rainfall appears to start around the end of the second decade of March, middle of the third decade of March, mid-April, end of the first decade of May, and early June, respectively, and end of the first decade of July, respectively. Also, rainfall at various stations appears to retreat, starting from the early third decade of October, the early third decade of October, end of the first decade of October, end of September, and early second decade of September, respectively end of the second decade of October, middle of the first decade of October, early October, and middle of the first decade of September respectively (Ngo 2019).

As a country, Nigeria has suffered significant loss because of the flood disaster, and the menace occurs annually. Nigeria has two major seasons: the rainy season and the dry season. A flood typically occurs at the peak of the rainy season, and the duration depends on the location (rainfall station). "It usually lasts (June-September) in the northern part and the remaining months are dry season with harmattan at (December-January); The southern part and the Nigerian delta areas experience rainfall between (March-October) with a short break in August; the peak of rainfall is in August and early September" (Ngo 2019). The record has it that more than 2.3 million people were displaced, 363 lost their lives, and about 16 million were affected in various ways. The effect has negated years of development in the country's economy, finance sector and caused human losses. Total losses were US$16.9 billion (Echendu 2020). Recent records have it that 155 died, and 25,000 were displaced after weeks of flooding between September and October 2020, as reported on October 20, 2020, according to flood list news in Africa by Copernicus Europe's eyes on Earth. The management agency cannot genuinely obtain the actual figure on the extent of fatality losses or displacement due to flooding in developing countries like Nigeria. Nigerian environments close to delta regions and stations prone to heavy rainfall are prone to the menace of a flood. The effect will slow down SDGs' progress. The negative impact is felt mainly through the people living in prone areas. Presently, there is no concrete research on finding the pattern of flooding in different flood zones or estimating what it will take to bring a lasting solution to the menace. For further reading on the impact of flooding in Nigeria, see (Echendu 2020). The outcome of these complications is that the collection system is "data-rich but information-poor." The procedure of our approach addresses this problem.

[16] stated that the problem could be formulated "mathematically" as follows: "Supposing there are n data variables (or features) and p sensor measurements at different recorded time instants, each measure representing an m-dimensional vector $x_i \in \mathbb{R}^n,\ i = 1, 2, \dots, p.$"; "The data are then assembled into a matrix $X = [, x_2, \dots, x_P]^T \in \mathbb{R}^{n \times P}$; each column denotes a process variable measured by one sensor operating alone" (Guo and Banerjee 2017). The significance of the method we proposed is that it automatically fixes threshold values for variables. It uses classifiers in binary classification and can combine with other techniques. It works well in large and noisy data via dimensionality reduction. Besides the fact that TDA and ML are recent research in statistics, the study will also help meet the standard for sustainable development goal (SDGs) related to flooding in Nigeria and help move the country toward economic growth, financial growth and minimize the loss of life. Our study used real data and, to the best of our knowledge, is the first to apply TDA in studying flooding in Nigeria.

2 Reasons for TDA and its fundamental ideas

A little close observation of the formula shows that x can represent a set of points in a data set, where x_i is a row vector defining sample i for which a lens value is calculated and x_j, are all the other samples in the data set. The output lens uses a row-wise Gaussian kernel estimator (Gaussian Density) over the data, and it is stated as

$$\boldsymbol{x_i} = \sum_{\boldsymbol{x_j} \in \boldsymbol{data\ set}} \boldsymbol{e^{-d^2(x_i, x_j)}} \tag{1}$$

According to (Offroy and Duponchel 2016), measuring shape and representation of the same are the two main tasks TDA performs; its important idea is that it considers a data set to be a point cloud (a sample) taken from a manifold in some high-dimensional space figure 1(i). (Offroy and Duponchel 2016) revealed the reason why topology is well suited for the analysis of big data sets in many areas. It was mentioned that simplices are constructed from the sample data, and those simplices develop intervals, which joined together to form a kind of wireframe approximately of the manifold (Offroy and Duponchel 2016). Three key properties that gave TDA the power for analyzing and understanding shape are (a) coordinate freeness; the principal idea says it should not matter so much how we represent the data in terms of coordinate provided we keep in track with the internal similarities (distances): that implies that the measure of topological shape does not change even if you rotate the shape or change the coordinate system of the shape view. The two

letters of A could constitute a set of data samples analyzed with two analytical forms (coordinate) while the topological construction extracts its main feature **figure 1**(ii). (b) deformation invariance. Topological properties are unchangeable even when there is a deformation or stretching of its; the letter A in **figure 1**(iii) is a loop with two legs and closed triangle, maintains the key features retrieved in the topological representation. This second property permits the measure of irregular or coarse data shape (holes, clumps, and voids), which are algebraically computed and quantitatively reflected. (c) compressed representation. If we observe a bit more closely into detail, a simple representation of the vital properties of the letter A (i.e., a close triangle and two legs can be observed) **figure 1**(iv). In consideration of the characteristic, the letter A is big data with millions of points of which TDA can generate a topological network with five nodes and five edges. These reasons make possible these highly scalable methods to produce a good opportunity to analyze immense (complex) data sets generated in various fields (medicine, computer science, biology, chemistry and so on) (Offroy and Duponchel 2016). (Ahmed **et** al. 2021) presents a random walk-based technique named EPD-RW that uses a topological structure of PPI (Protein-Protein) to identify essential proteins and integrated network topology and biological information extracted from GO (gene ontology) data, gene expression profiles, domain information and phylogenetic profile.

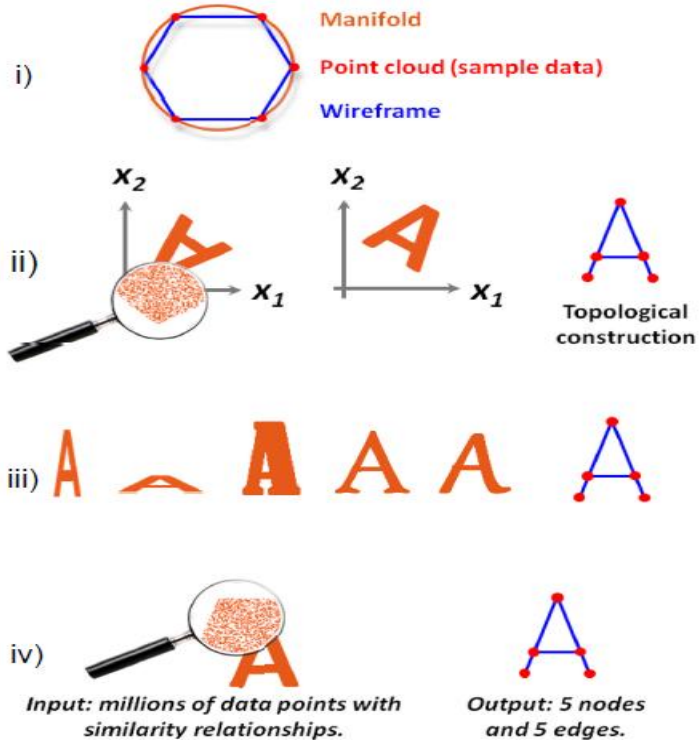

Fig. 1 (i) Vital idea of TDA. The three key properties: (ii) coordinate invariance, (iii) deformation invariance and (iv) compression representation

2.1 The general framework of topological data analysis

(Offroy and Duponchel 2016) explained the steps in the formulation of TDA: In the first step, called partition, a metric and a lens are chosen. The metric are the measure of distance or similarity between any two data points while the lens (circular file) is the mathematical function through which data are viewed or observed. The most important point here is that anything, producing a number from a data point can be a lens and gives us the freedom to choose different functions that provide different viewpoints. Lenses can come from statistics (mean, maximum, minimum, …), from geometry (centrality, curvature, …), from chemometrics (PCA scores, SVM Distance from hyperplane, MDS scores, …) and so on; the lens derives the division of data points into subpopulation (overlapping circular files) (Offroy and Duponchel 2016). A set can be observed simultaneously through different lenses by simply multiplying their effects. In step two (cluster analysis), partition is analyzed, and data are clustered within these bins in such a way that a cluster contains rows that are resembling each other. A row can be a spectrum characterizing a sample from a spectroscopic perspective. Because the data set is divided into bins (circular files) in an overlapping way, each row is oversampled and falls into more than one cluster (partition). In the third step, called network generation, data are reassembled to generate the final network. If two clusters in different bins share one or more rows, an edge is used between the two clusters to form the final network **figure 2**.

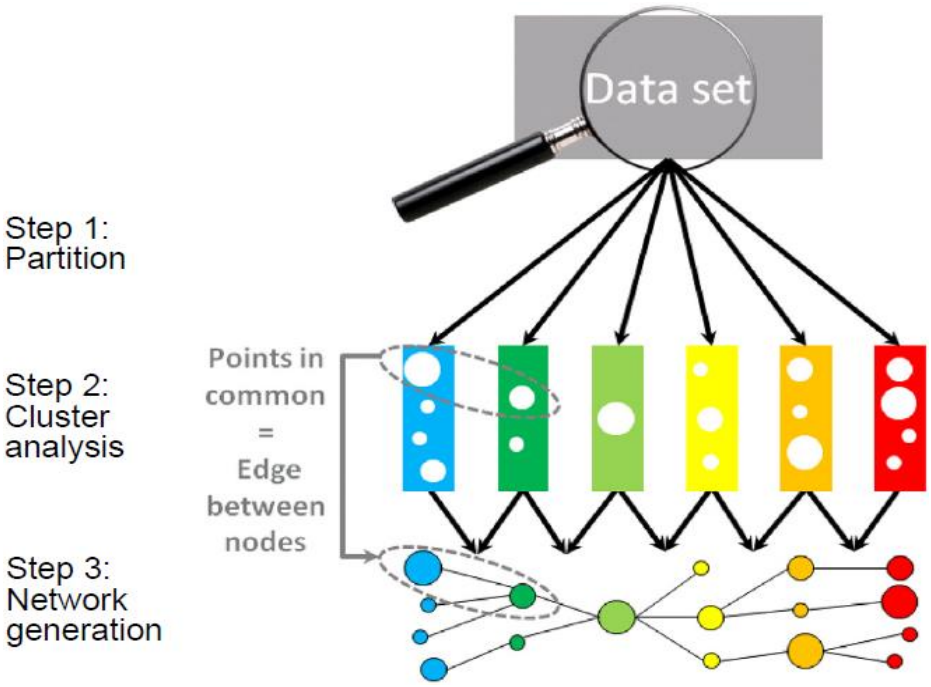

Fig. 2 Visualization of the general framework of TDA

Source: (Offroy and Duponchel 2016).

3. Materials and Methods

3.1 Studied Areas and Materials

We collected spatial data on four weather parameters (Maximum Temperature, Minimum Temperature, Rainfall & Windspeed) for seven states from 1970 to 2018 from the Nigeria Hydrological Services Agency (NIHSA). We have a total of 17 394 data points. The NIHSA, to the best of our knowledge, is the only agency in Nigeria that has a 48-year dataset. These states (Anambra, Bayelsa, Benue, Edo, Kogi, Kwara, and River) were considered based on the flooding information, as shown in **figure 3**. This research article uses quantitative analysis. It is measured using numbers and values with units in millimeter, kelvin, meter/second for precipitation, temperature, and wind speed.

Pattern recognition is usually categorized depending on the learning procedure that determines how the output result will be. The learning procedure is an unsupervised type (i.e., given a shape in a new group without labeled data). We can refer to the form as time-series trends in this study since our data point is in indexed time order. The procedure starts with:

1) Data collection; the data are quantitative and well-structured in a spreadsheet. We applied officially by first sending a mail to the Director of the NIHSA, a request for data collection. Attached to the application was a cover letter endorsed by our Institution's Assistant Registrar (Institute of Postgraduate Studies, USM). We received the data through email after about five months of applying to the agency. We then converted our data from daily data to monthly data using excel. As the data was being plotted in excel, we observed some outliers in the information on two dates (1/1/2008 and 24/2/2013).

2) Interpolation method; we used the Interpolation method to estimate observed missing values and outliers. We used font color to show the cell locations where interpolation is used in our data file with the interpolation formula. Interpolation reduces the effect of error in measurement and outliers in the data.

3) Loading data; we then loaded our data and continued with the automated method (TDA) using Python software codes. The procedure of our approach is two-fold. Stage 1 will apply topology and algorithms from TDA to compute topological features at different spatial resolutions. TDA uses information from all parameter values: it encodes them in a representable diagram (barcode).

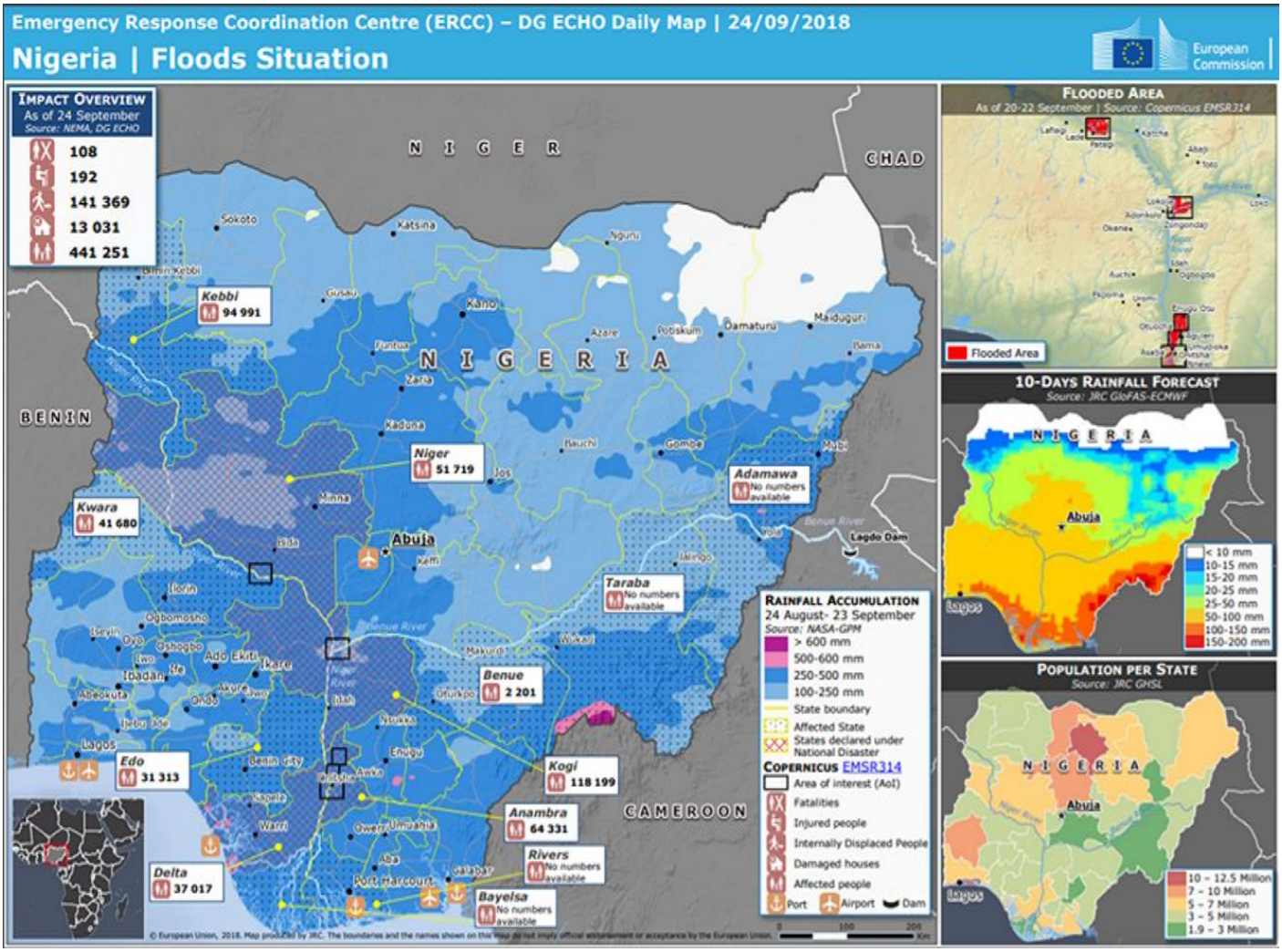

Fig. 3 Flood Areas in Nigeria. "European Commission's Directorate-General for European Civil Protection and Humanitarian Aid Operations (DG ECHO). (2018, September 24)**".**[1]

3.2 K-mean algorithm and Silhouette analysis

K-mean is an iterative clustering algorithm that works in the following five steps:

Step 1. Libraries' Importation: Importation of pandas as PD, np, plt, then K-mean from sklearn cluster; these are encoded in the computation spreadsheet.

Step 2. Importation of data: We use code to import data from excel, and we already saved each state with a different file name.

Step 3. Computation of the cluster centroids: The code for computing the centroid of data points is Kmean.cluster_centers.

[1] Source: (European Commission's Directorate-General for European Civil Protection and Humanitarian Aid Operations (DG ECHO) 2018, September 24).

Step 4. Testing of the algorithm: The code for getting the data points' labels are categorized into two clusters, namely, k-mean and k-medoids, both partitional clustering. We used k-mean and Silhouette analysis to test it.

Step 5. Re-computation: Re-computation of the cluster until a global optimal is achieved; else, repeat steps 4 and 5. K-mean Clustering and Silhouette analysis are implemented using R. The implementation procedure is clearly shown in **figure 4.**

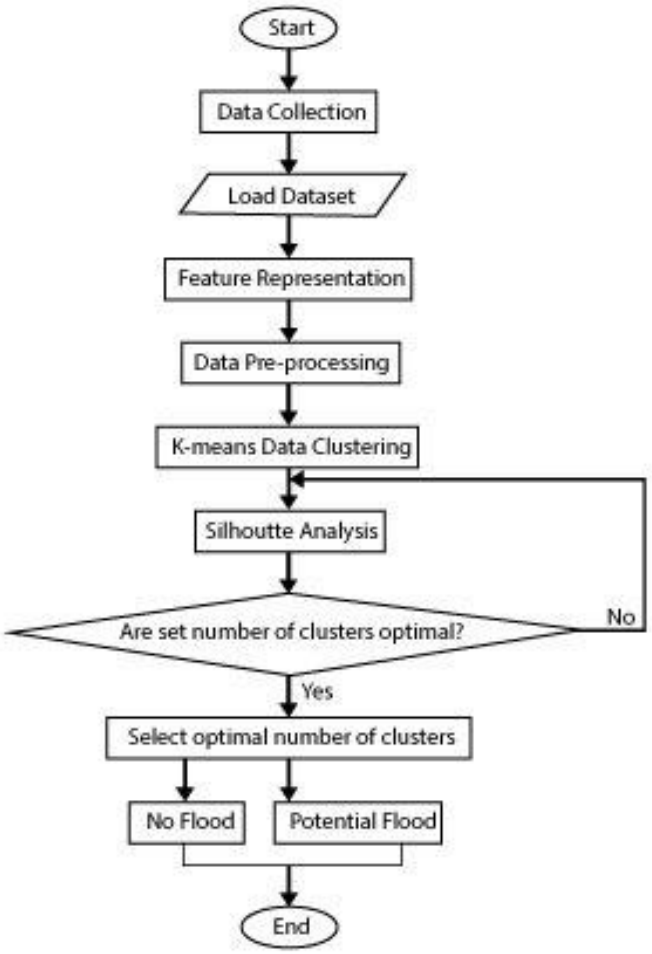

Fig. 4. Flow chart of Implementation Process

3.3 Methods in TDA

Clustering techniques or density clustering, persistent homology, and estimation in the manifold are TDA methods, all of which study feature (s) in TDA methodology. Clustering is the simplest version of TDA.

The equations that contributed to the study are: Let $X_1, ..., X_n$ be observation from a distribution W; $w =$ distribution′s density, $X_i \in \mathcal{X} \subset \mathbb{R}^d$.

$$X_1, \dots, X_n \sim W \tag{2}$$

W supported some set $\mathcal{X} \subset \mathbb{R}^d$. (Hartigan 1981) has it "For any $t \geq 0$ define the upper-level set."

$$L_t = \{x: w(x) > t\}. \tag{3}$$

"The density clusters at level t denoted by C_t are the connected component of L_tThe set of all the density clusters is"

$$C = \bigcup_{t \geq 0} C_t. \tag{4}$$

The estimated level set is

$$\hat{L}_t = \{x: \hat{w}(x) > t\} \tag{5}$$

"$\hat{w}$, is any density estimator of which a common choice is the kernel density estimator."

$$v(x) = \frac{1}{h}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{\|x - X_i\|}{h}\right) \tag{6}$$

"Where $h > 0$ is the bandwidth, and K is the kernel" (Wasserman 2018). "For the theoretical properties of the estimator," $\hat{L}_t$ see (Rinaldo and Wasserman 2010) and (Cadre 2006).

3.4 Cluster Validity Indices

In applying the Validity test, we used R software. We first installed some useful R packages into the R library to enable us to implement coded functions. We saved our data in a .csv file, which R software can read via a code, read.csv ("Felix data cluster.csv"). We then scaled/standardized our data using the R function scale to ensure a constant variable unit. We do not want the clustering algorithm to depend on an arbitrary variable unit (metric). We then applied codes in R to carry out the validity of our proposed method.

There are various performance measures to help in assessing validity indices. Validity indices entail the following measurement (comparison):

- Comparison of clustering algorithms
- Comparison of two sets of groups within connectedness
- Determination of a variable dimensionality due to noise within the dataset

(Handl, Knowles, and Kell 2005) Stated that, "A good clustering algorithm generates clusters with high intra-cluster homogeneity, good inter-cluster separation and high connectedness between close data points." (Handl, Knowles, and Kell 2005) further mentioned that, "The visualization of clusters has several advantages over traditional performance curves; it allows one to summarize information regarding the algorithms' performance under both internal validity measures." (Brock et al. 2011) pointed out that, "the validity measures

are generally categorized into three classes of internal, stability and biological." Our research used the internal validity measurement that includes the Dunn test, Silhouette, and Gap statistics.

3.4.1 Dunn's index

A British army officer, James C. Dunn, introduced the Dunn index (DI) in 1974, as reported by (Dunn 1974). It is a metric for examining clustering algorithms. It belongs to the kind of measurement embedded with internal information. Dunn's test can measure compactness in the cluster. It is indicated where the mean of different groups (collection) is sufficiently farther apart than inside the set (group). It is recorded that when Dunn's value is more significant, its configuration performance is more superior. The number of clusters k that maximizes Dunn's test is chosen as the optimal cluster k**.** The Dunn's test has some constraints; As the number of groups and dimensionality of the data increase, the computational cost also increases (Dunn 1974). The formulation of Dunn's index for k clustered value is shown in **table 1**.

3.4.2 Silhouette coefficient (SC)

Silhouette is a metric index used to interpret and validate consistency within clusters of data [30]. The technique provides an exact graphical representation of how well each object has been classified. The Silhouette value also measures how similar an item is to its cluster (closeness) compared to other groups (separation). The Silhouette's range is between (−1 to +1), and if most objects have a high value (i.e., range 0,1), then the clustering configuration (output) is fitting. If many points are of low or negative value, then the clustering feature may have too many or too few clusters. "An improved version of the Silhouette coefficient (SC) minimizes the computation time on distance calculations by reducing the number of addition operations" (Zhao 2012).

3.4.3 Gap statistic (GS)

"R. Tibshirani, G. Walther, and T. Hastie (Sandford University, 2001) published Gap statistic, and the approach was applied to any clustering method (i.e., K-mean clustering, hierarchical clustering)"; "it compares the total intra-cluster variation for different values of *k* with their expected values under null reference distribution of the data" (Tibshirani, Walther, and Hastie 2001). The validity test standardized the graph under an appropriate null reference distribution of the data (Tibshirani, Walther, and Hastie 2001). Gap statistic results are almost the same compared with our work using python, which shows high efficiency.

3.4.4 The Classical Metric Distance for Calculating Cluster Validity index

The Euclidean distance or the Manhattan distance is an efficient distance measure used in calculating **SC** and **DI**. The distance (space) calculation is referred to as a metric if it satisfies three properties: (i) symmetry, (ii) triangle inequality, and (iii) positive value.

Euclidean $\boldsymbol{x_i = (x_{i1},\ x_{i2},\ \ldots, x_{ip})}$ and $y_j = (\boldsymbol{y_{j1},\ y_{j2},\ \ldots, y_{jp}})$

$$\boldsymbol{d(x_i, y_j) = \left(\left|x_{i1} - y_{j1}\right|^2 + \left|x_{i2} - y_{j2}\right|^2 + \left|x_{i3} - y_{j3}\right|^2 + \ldots + \left|x_{ip} - y_{jp}\right|^2\right)^{\frac{1}{2}}}$$

$$\boldsymbol{d(x_i, y_j) = \left(\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_{i1} - y_{j1}\right|^2\right)^{\frac{1}{2}} \quad \forall x_i,\ y_j \in S}$$

$$\boldsymbol{d_{\text{man}}(x_i, y_j) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}} \tag{7}$$

Manhattan $\boldsymbol{x_i = (x_1, x_2, \ldots x_n)}$ **and a point** $\boldsymbol{y_j = (y_1, y_2, \ldots y_n)}$ **is:**

$$\boldsymbol{d_{\text{man}}(x_i, y_j) = \sum_{i=1}^{n}|(x_i - y_i)|} \tag{8}$$

n, is the number of variables, and x_i and y_i are the values of the i^{th} variable, at points x and y, respectively.

Some of the significant factors affecting specific validation techniques are based on two-dimensional datasets "Long and Square" (Rousseeuw 1987, Handl, Knowles, and Kell 2005). For this, advanced techniques are selected among the following F-measure (takes values in [0,1], to be maximized), the adjusted Rand index (takes points in [0,1], to be maximized), variance (to be minimized), connectivity (to be minimized), Silhouette Width ([−1,1], to be maximized), the Dunn Index (to be maximized) and a stability based method (i.e., in [0,1], to be maximized).

Table 1: Formulation of Internal Validity

Validation	*Formulation*	*Value*
Dunn index (DI_n)	$DI_n = \frac{\min\limits_{1\le i<j\le n} dist(c_i, c_j)}{\max\limits_{1\le i\le n} diam(c_l)}$ (9) $dist(c_i, c_j)$ = inter-cluster distance metric between c_i and c_j clusters; $diam(c_l)$ Calculates the "maximum distance," "it is a distance of all points from the mean" and, "the mean distance between all the pairs."	Between 0 and infinity (∞). The index of Dunn should be maximized (large).
Silhouette width	$S_i = \frac{b_i - a_i}{max\{a_i, b_i\}}$ (10) $b_i = min \frac{\sum_j d(i,j)}{\lvert c_j \rvert},\ C_i \neq C_j$ $a_i = \min \frac{\sum_j d(i,j)}{\lvert c_j \rvert},\ C_i = C_j$ a_i = average distance between i & j all other observations in the same cluster *C* b_i = average distance between i & j all other observations in the nearest	Between range $(-1 \& +1)$. The Silhouette with positive values are satisfactory, and values close to +1 are perfect measures, but those with negative values are unwell observations.
Gap statistic	$Gap_n(K) = E_n^*\{log(W_k)\} - log(W_k)$, (11) $D_h = \sum_{i,j\in C_h} d_{i,j}$ *and* $W_k = \sum_{h=1}^{k} \frac{1}{2n_h} D_h$ E_n^* = the expectation under the sample of size n in the distribution D_h= the pairwise distances for all the points in cluster *h,* *d* = squared Euclidean space, and the set W_k= pooled within-cluster sum of squares around the mean cluster (factor 2 makes it exactly).	A reference distribution is given by uniform distribution, U [0,1].

3.4.5 The algorithm's procedure for the implementation of cluster validity (Optimal Validity)

The algorithm implementation procedure in k-mean using the R software code is shown in **figure 5.**

Procedure in K-Means

Set $\vec{\mu}_i, \ldots, \vec{\mu}_k$ to be distinct randomly selected inputs from $\vec{x}_1, \ldots, \vec{x}_n$

repeat

for i = 1 ... n **do** // fix $\vec{x}_1, \ldots, \vec{x}_n$ update γ

$$\gamma_{ij} = \begin{cases} 1 \text{ if } j = \arg\min_j \left((\vec{x}_j - \vec{\mu}_j)^{\frac{1}{2}} \right)^2 \text{ for Euclidean and } \arg\min_j |\vec{x}_j - \vec{\mu}_j| \text{ for Manhattan} \\ 0 \text{ otherwise} \end{cases}$$ // assign cluster membership

end for

for j = 1 ... k **do** // fix γ, update $\vec{\mu}_i, \ldots, \vec{\mu}_k$

$n_j = \sum_{i=1}^{n} \gamma_{ij}$ // # of points assigned to cluster j

$\vec{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n} \gamma_{ij} \vec{x}_i$ // $\vec{\mu}_j$ is the set to the average of all points assigned to cluster j

end for

until convergence

end procedure

return $\vec{\mu}_i, \ldots, \vec{\mu}_k$

Fig. 5. k-mean Algorithms

4 Results and Discussion

The results of the seven states were presented, with their patterns based on the real data from Nigeria Hydrological Services Agency. The findings show potential flood and no flood zones respectively for each region under study. The resultant Silhouette analysis of data points in the generated clusters is shown in **Table 2**. The green and black dots in the obtained groups represent different patterns of partitioning. Those patterns-belonging to 0 groups are non-AR, which indicates no flooding, while others belonging to 1 are the AR, meaning potential-flooding **figure 8**. The dotted red line in **figure 7** indicates the Silhouette coefficient (at k = 2) in the plot. The Densities of features in the dataset on the four parameters were obtained for each state, and the summary is shown in **figure 9**. The average points generated in the Silhouette Analysis are 0.5748, 0.6332, 0.5200, 0.5216, 0.5002, and 0.4790 for Anambra, Bayelsa, Benue, Edo, Kogi, and Kwara, respectively; see **Table 2**. At k = 2, the Silhouette coefficient's range lies between 0.7 and 0.8 (70% to 80%), which measures how excellent and efficient our analysis is. And the performance is expected to improve if more parameters were available. The values of the variances 0.0205563, 0.0132712, 0.0203055, 0.0308588,

0.0239802, 0.0284630, and, 0.0085098 obtained for each state are small, which shows that the clusters in each set are not spread apart.

The calculated variance indicated that the average squared distance is short, which implies no widespread within the clusters. We conducted a validity test on our proposed technique in R software codes to evaluate the performance. In the cluster validity, three tests (Dunn's index (DI), Silhouette (SW) and Gap statistic (GS)) were used; the aim is to validate the goodness of fit of our analysis and to compare the clustering algorithm that best measures our method at cluster k = 2. The three-validity test performed very well considering the range of their outcomes result. The GS and the SI fall on almost the same range value; the three-evaluation test also follows the same time-series trend (pattern) considering their values. The optimal values, produced in the validity tests at k = 2, fall within the same range as in the result we obtained using Python; validity tests fall between 0.6 and 0.8 (60% to 80%) at k = 2, which shows that our method is highly efficient. **Figure 10** in the validity test produced better clustering configuration and better connectedness in the partitions in comparison to clustering algorithms. The SI plot in **figure 11** indicated the optimum value at exactly k = 2 in all seven states.

Table 2: Summary of Silhouette Analysis of Data Points in Generated Clusters

State	***Number of clusters (k)***					***Average***	***Variance***
	2	**3**	**4**	**5**	**6**		**s^2**
Anambra	0.722	0.660	0.599	0.539	0.348	0.5736	0.0205563
Bayelsa	0.788	0.700	0.630	0.544	0.504	0.6332	0.0132712
Benue	0.697	0.602	0.548	0.352	0.401	0.5200	0.0203055
Edo	0.708	0.634	0.530	0.488	0.248	o.5216	0.0308588
Kogi	0.700	0.600	0.504	0.349	0.348	0.5002	0.0239802
Kwara	0.710	0.608	0.344	0.354	0.379	0.4790	0.0284630
River	0.759	0.683	0.627	0.551	0.527	0.6294	0.0085098
Range	(0.7–0.8)	(0.6–0.7)	(0.3–0.6)	(0.3–0.5)	(0.2–0.5)	***Average range*** **(0.3–0.8)**	

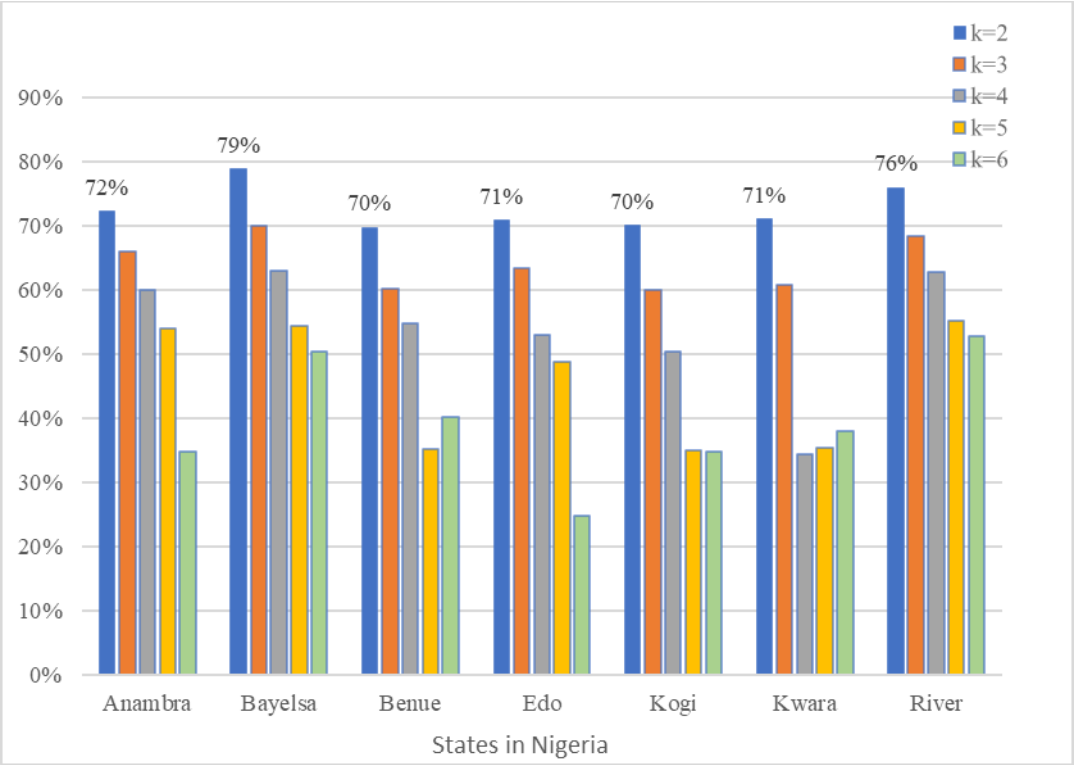

Fig. 6 The Summary of Silhouette Analysis Score on Different Number of Cluster

FIGURE 7 HERE

Fig. 7 The Summary of Silhouette Analysis Score with the red line indicating score at k = 2

FIGURE 8 HERE

Fig. 8 The summary of clustered feature pattern on each of the seven state

FIGURE 9 HERE

Fig. 9 The summary of the density of features on the four variables parameters in each state.[2]

4.1 Validity plots of the clusters

The cluster validity test of our proposed technique conducted in R software codes is used to evaluate the cluster's performance of our analysis. Table 3 shows the values obtained from the three tests (Dunn's index (DI), Silhouette (SI), and Gap statistic (GS)). **Figures 10, 11,** and **12** are the plots of the validity measurement

[2] The density plot in **figure 9** is the actual resultant density of the four variable parameters (Maximum and Minimum Temperature (K), Precipitation (mm), Wind Speed (m/s)) used on each of the seven states.

we conducted. The SI plot in **figure 11** best evaluated our study based on the displayed (observed) characteristic. Close observation shows that SI plots attained the optimality at k = 2, thus evaluating our analysis at exactly k = 2 in all the states. Better connectivity in the partitions was achieved and measured how accurate our analysis is. In **figure 13a** and **13b**, the SW and GS patterns are the same and are represented with A; the pattern for DI is illustrated with B: Close observation on **figure 13a** and **13b** show that plots A and plots B follow the same feature patterns in each of seven states, irrespective of the shape and, the use of different codes in R software. This fundamental property shows that the three cluster validity tests have some features they share and thereby validated the three tests as authentic.

TABLE 3 HERE

Table 3 Summary of the cluster validity test at k = 2

FIGURE 10 HERE

Fig. 10 The summary of the cluster validity feature pattern on each of the seven states.[3]

FIGURE 11 HERE

Fig. 11 The visualization of Silhouette's optimal plot

FIGURE 12 HERE

Fig. 12 The visualization of the Gap statistics plot

FIGURE 13a HERE

Fig. 13 Summary plots of the cluster validity feature pattern on four states

FIGURE 13b HERE

Fig. 13 Summary plots of the cluster validity feature pattern on three states

[3] In comparison of clustering algorithms, the figure in the validity test produced better clustering configuration and better connectedness in the dataset.

4.2 The variance the Clusters

A variance for a set of measurement is the average squared distance from the mean. When a finite set of data or population is being used, the mean square deviation is called population variance. Given a set of data $x_1, \ldots, x_n$, the formulation is given as follows.

$$\sigma_x^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \quad n < 30 \text{ or } \sigma_x^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} \quad n \geq 30$$

(11)

In this study, x is given in table 4.31 and $n = 5$. The calculation is done using Excel Microsoft given as shown below.

Table 4 Variance of the Clusters for the 7 selected States

1) Anambra State			
Count	**Obsn(n)**	**x-mean**	**(x-mean)^2**
2	0.722	0.1484	0.0220226
3	0.66	0.0864	0.007465
4	0.599	0.0254	0.0006452
5	0.539	-0.0346	0.0011972
6	0.348	-0.2256	0.0508954
Sum	2.868	0	0.0822252
Count(n)	5	5	5
Ave(mean)	0.5736		
Var			0.0205563
SD			**0.1433747**
2) Bayelsa State			
Count	**Obsn(n)**	**x-mean**	**(x-mean)^2**
2	0.788	0.1548	0.023963
3	0.7	0.0668	0.0044622
4	0.63	-0.0032	1.024E-05
5	0.544	-0.0892	0.0079566
6	0.504	-0.1292	0.0166926
Sum	3.166	0	0.0530848
Count(n)	5	5	5
Ave(mean)	0.6332		
Var			0.0132712
SD			

3) Benue State			
Count	**Obsn(n)**	**x-mean**	**(x-mean)^2**
2	0.697	0.177	0.031329
3	0.602	0.082	0.006724
4	0.548	0.028	0.000784
5	0.352	-0.168	0.028224
6	0.401	-0.119	0.014161
Sum	2.6	4.44089E-16	0.081222
Count(n)	5	5	5
Ave(mean)	0.52		
Var			**0.0203055**
SD			
4) Edo State			
Count	**Obsn(n)**	**x-mean**	**(x-mean)^2**
2	0.708	0.1864	0.034745
3	0.634	0.1124	0.0126338
4	0.53	0.0084	7.056E-05
5	0.488	-0.0336	0.001129
6	0.248	-0.2736	0.074857
Sum	2.608	0	0.1234352
Count(n)	5	5	5
Ave(mean)	0.5216		
Var			**0.0308588**
SD			
5) Kogi State			
Count	**Obsn(n)**	**x-mean**	**(x-mean)^2**
2	0.7	0.1998	0.03992
3	0.6	0.0998	0.00996
4	0.504	0.0038	1.444E-05
5	0.349	-0.1512	0.0228614
6	0.348	-0.1522	0.0231648
Sum	2.501	5.55112E-16	0.0959208
Count(n)	5	5	5
Ave(mean)	0.5002		
Var			**0.0239802**

SD			
6) Kwara State			
Count	**Obsn(n)**	**x-mean**	**(x-mean)^2**
2	0.71	0.231	0.053361
3	0.608	0.129	0.016641
4	0.344	-0.135	0.018225
5	0.354	-0.125	0.015625
6	0.379	-0.1	0.01
Sum	2.395	0	0.113852
Count(n)	5	5	5
Ave(mean)	0.479		
Var			0.028463
SD			
7) River State			
Count	**Obsn(n)**	**x-mean**	**(x-mean)^2**
2	0.75	0.1224	0.0149818
3	0.683	0.0554	0.0030692
4	0.627	-0.0006	3.6E-07
5	0.551	-0.0766	0.0058676
6	0.527	-0.1006	0.0101204
Sum	3.138	-1.11022E-16	0.0340392
Count(n)	5	5	5
Ave(mean)	0.6276		
Var			**0.0085098**
SD			

5. Conclusions

T TDA is useful for data with high order interaction. The visualization of the clustered data for each state studied the “no flooding” areas and the “potential flooding” zones. The pattern obtained for each flood zone is useful in predicting flooding, no flooding, and identifying zones with the highest flood rate in the considered regions. The average Silhouette values of the data are criteria used for assessing the natural number of clusters. Silhouette coefficients near one gave an excellent fit to the group dataset (i.e., the ideal set), but values relative to 0 were unsatisfactory clusters; negative values indicate outliers in the set. The Silhouette results obtained

in our study lie within the range (0.3 and 0.8) for all seven states that are considered, and that implies that there is no wrong cluster in our work. At k = 2, the result fell between 0.7 and 0.8 (70% to 80%), which means that our analysis is efficient. After our model's training process, we obtained the best group at k = 2, where we have the highest Silhouette coefficient for each of the seven states. That gives us an efficiency outcome of approximately 80%, which is expected to improve with the increase in data points. The TDA method's data pattern is displayed in the visualization of clustered data for every seven states; the bar chart in **figure 6** summarizes the score on different cluster numbers. The result (pattern obtained) can also predict imminent danger and inform the people living in the flood region to vacate from the zone and avert havoc due to flood. The performance of the approach used in the study has been evaluated for accuracy and precision; the intra-cluster validity test conducted showed higher evaluation measurement. The three validity tests have performed very well based on their outcome; we chose the Silhouette test as the best validity test based on the outcome plot in **figure 11**; it indicated the optimum value at exactly k = 2 in all seven states. With that, our objective in comparison has been achieved. The variance is not widespread in the cluster, which fulfilled our aim to evaluate the extent of cluster variability or spread. The version of the study is expected to improve if more rainfall parameters were available and used. It has been reviewed in Section 1 that an increase in data points will increase the performance (Riihimäki et al. 2020). Our method takes care of a square error criterion by using unsupervised learning integrated into the TDA technique. Some of the highlighted problems common to ordinary methods have been addressed automatically in our study. Most of the literature that adopted the same approach showed that our method scores are high in classification accuracy and pattern detection. From the conducted validity test, high homogeneity was shown in connectedness at k=2 **figure 10**. This study discovered that there existed a common relationship among the three validity tests in terms of similarity in feature patterns figures **13a** and **13b**. Future research should be focused on using more variables on flooding for the study. We recommend further study on the combination of our method with other methods since our approach is flexible. The limitations encountered on the course of the research are

1) Lack of access to a database.

2) The unavailability of sufficient weather parameters: only four weather parameters were available during data collection.

3) Time: The total period between sending out the application and receiving the data took about five months.

4) Maintenance: We witnessed a lack of adequate maintenance and replacement of some faulty machines and malfunctioning sensors.

Funding. The authors sincerely declare that there is no funding for this research.

The availability of data. The datasets generated and analyzed during this study are not publicly available due (due to the collection processes that involved application, declaration of the data range needed, payment of money before the agency (NIHSA) would release the data by mail) but are available from the corresponding author on reasonable request.

Supplement. The supplement related to this article is not available online. It is collected only through official request and application to the email (info@nihsa.gov.ng) of the NIHSA.

Competing interest. The authors declared that they have no conflict interests.

Disclaimer. This article was prepared as an account of research work done at the Universiti Sains Malaysia. This document is believed to contain the correct information, neither the Federal Government of Nigeria nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, or process disclosed, or represents that its use would not infringe privately owned rights. Reference here to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favor by the Nigerian Government or any agency thereof. The authors' views and opinions expressed here do not necessarily state or reflect those of the Nigerian Government or any agency thereof.

.

Acknowledgment

Words of appreciation go to the Universiti Sains Malaysia for the training and support.

References

Abou El Majd, B., Driss Bennis, Fouad Gharib, Ghita Lebbar, and H. El Ghazi. 2018. "Persistent Homology applied to location problems." *MATEC Web of Conferences* no. 200:00003. doi: 10.1051/matecconf/201820000003.

Ahmed, Nahla Mohamed, Ling Chen, Bin Li, Wei Liu, and Caiyan Dai. 2021. "A random walk-based method for detecting essential proteins by integrating the topological and biological features of PPI network." *Soft Computing*. doi: 10.1007/s00500-021-05780-8.

Alaa, H. N., and S. A. Mohamed. 2017. "On the Topological Data Analysis extensions and comparisons." *Journal of the Egyptian Mathematical Society* no. 25 (4):406-413. doi: 10.1016/j.joems.2017.07.001.

Brock, Guy, Vasyl Pihur, Susmita Datta, and Somnath Datta. 2011. "clValid, an R package for cluster validation." *Journal of Statistical Software (Brock et al., March 2008)*.

Bubenik, Peter. 2015. "Statistical topological data analysis using persistence landscapes." *The Journal of Machine Learning Research* no. 16 (1):77-102.

Cadre, Benoıˆt. 2006. "Kernel estimation of density level sets." *Journal of multivariate analysis* no. 97 (4):999-1023.

Carlsson, Gunnar, Afra Zomorodian, Anne Collins, and Leonidas J Guibas. 2005. "Persistence barcodes for shapes." *International Journal of Shape Modeling* no. 11 (02):149-187.

Chevyrev, I., V. Nanda, and H. Oberhauser. 2018. "Persistence paths and signature features in topological data analysis." *IEEE transactions on pattern analysis and machine intelligence*.

de Gois, Givanildo, José Francisco de Oliveira-Júnior, Carlos Antonio da Silva Junior, Bruno Serafini Sobral, Paulo Miguel de Bodas Terassi, and Antonio Herbete Sousa Leonel Junior. 2020. "Statistical normality and homogeneity of a 71-year rainfall dataset for the state of Rio de Janeiro—Brazil." *Theoretical and Applied Climatology* no. 141 (3):1573-1591.

Dunn, Joseph C. 1974. "Well-separated clusters and optimal fuzzy partitions." *Journal of cybernetics* no. 4 (1):95-104.

Echendu, Adaku Jane. 2020. "The impact of flooding on Nigeria's sustainable development goals (SDGs)." *Ecosystem Health and Sustainability* no. 6 (1):1791735.

Edelsbrunner, Herbert. 2013. "Persistent homology: theory and practice."

European Commission's Directorate-General for European Civil Protection and Humanitarian Aid Operations (DG ECHO). *Nigeria | Floods Situation - Emergency Response Coordination Centre (ERCC)* 2018, September 24.

Available from https://reliefweb.int/map/nigeria/nigeria-floods-situation-emergency-response-coordination-centre-ercc-dg-echo-daily-map.

Farrelly, Colleen Molloy. 2017. Topological Data Analysis for Data Mining Small Educational Samples with Application to Studies of the Gifted.

Frosini, Patrizio. 1990. "A distance for similarity classes of submanifolds of a Euclidean space." *Bulletin of the Australian Mathematical Society* no. 42 (3):407-415.

Gholizadeh, Shafie, and Wlodek Zadrozny. 2018. "A Short Survey of Topological Data Analysis in Time Series and Systems Analysis." *arXiv preprint arXiv:1809.10745*.

Gidea, Marian, and Yuri Katz. 2018. "Topological data analysis of financial time series: Landscapes of crashes." *Physica A: Statistical Mechanics and its Applications* no. 491:820-834.

Guiang, Chona S, and Robert Y Levine. 2012. Cloud detection and characterization using topological data analysis. Paper read at Remote Sensing of Clouds and the Atmosphere XVII; and Lidar Technologies, Techniques, and Measurements for Atmospheric Remote Sensing VIII.

Guo, Wei, and Ashis G Banerjee. 2017. "Identification of key features using topological data analysis for accurate prediction of manufacturing system outputs." *Journal of Manufacturing Systems* no. 43:225-234.

Handl, Julia, Joshua Knowles, and Douglas B Kell. 2005. "Computational cluster validation in post-genomic data analysis." *Bioinformatics* no. 21 (15):3201-3212.

Hartigan, John A. 1981. "Consistency of single linkage for high-density clusters." *Journal of the American Statistical Association* no. 76 (374):388-394.

Khasawneh, Firas A, Elizabeth Munch, and Jose A Perea. 2018. "Chatter classification in turning using machine learning and topological data analysis." *IFAC-PapersOnLine* no. 51 (14):195-200.

Krim, Hamid, Thanos Gentimis, and Harish Chintakunta. 2016. "Discovering the Whole by the Coarse: A topological paradigm for data analysis." *IEEE Signal Processing Magazine* no. 33 (2):95-104. doi: 10.1109/msp.2015.2510703.

Letscher, H Edelsbrunner D, and A Zomorodian. 2002. "Topological Persistence and Simplification." *Discrete Computational Geometry* no. 28:511-533.

Muszynski, Grzegorz, Karthik Kashinath, Vitaliy Kurlin, and Michael Wehner. 2019. "Topological data analysis and machine learning for recognizing atmospheric river patterns in large climate datasets." *Geoscientific Model Development* no. 12 (2):613-628.

Ngo, MBE. 2019. "Human Rights Issues in Cameroon in the Case of the Independentists Arrested in Nigeria and Extradited to Cameroon." *Human Rights Issues in Cameroon in the Case of the Independentists Arrested in Nigeria and Extradited to Cameroon (October 21, 2019)*.

Offroy, Marc, and Ludovic Duponchel. 2016. "Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry." *Analytica chimica acta* no. 910:1-11.

Pascucci, Valerio, Xavier Tricoche, Hans Hagen, and Julien Tierny. 2010. *Topological Methods in Data Analysis and Visualization: Theory, Algorithms, and Applications*: Springer Science & Business Media.

Richeson, David S. 2019. *Euler's gem: the polyhedron formula and the birth of topology*: Princeton University Press.

Riihimäki, Henri, Wojciech Chachólski, Jakob Theorell, Jan Hillert, and Ryan Ramanujam. 2020. "A topological data analysis based classification method for multiple measurements." *BMC bioinformatics* no. 21 (1):1-18.

Rinaldo, Alessandro, and Larry Wasserman. 2010. "Generalized density clustering." *The Annals of Statistics* no. 38 (5):2678-2722.

Robins, Vanessa. 1999. Towards computing homology from finite approximations. Paper read at Topology proceedings.

Rousseeuw, Peter J. 1987. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics* no. 20:53-65.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie. 2001. "Estimating the number of clusters in a data set via the gap statistic." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* no. 63 (2):411-423.

Umeda, Yuhei, Junji Kaneko, and Hideyuki Kikuchi. 2019. "Topological Data Analysis and Its Application to Time-Series Data Analysis." *FUJITSU SCIENTIFIC & TECHNICAL JOURNAL* no. 55 (2):65-71.

Wasserman, Larry. 2018. "Topological data analysis." *Annual Review of Statistics and Its Application* no. 5:501-532.

Zhao, Qinpei. 2012. "Cluster validity in clustering methods." *Publications of the University of Eastern Finland*:16.

Zomorodian, Afra, and Gunnar Carlsson. 2005. "Computing persistent homology." *Discrete & Computational Geometry* no. 33 (2):249-274.

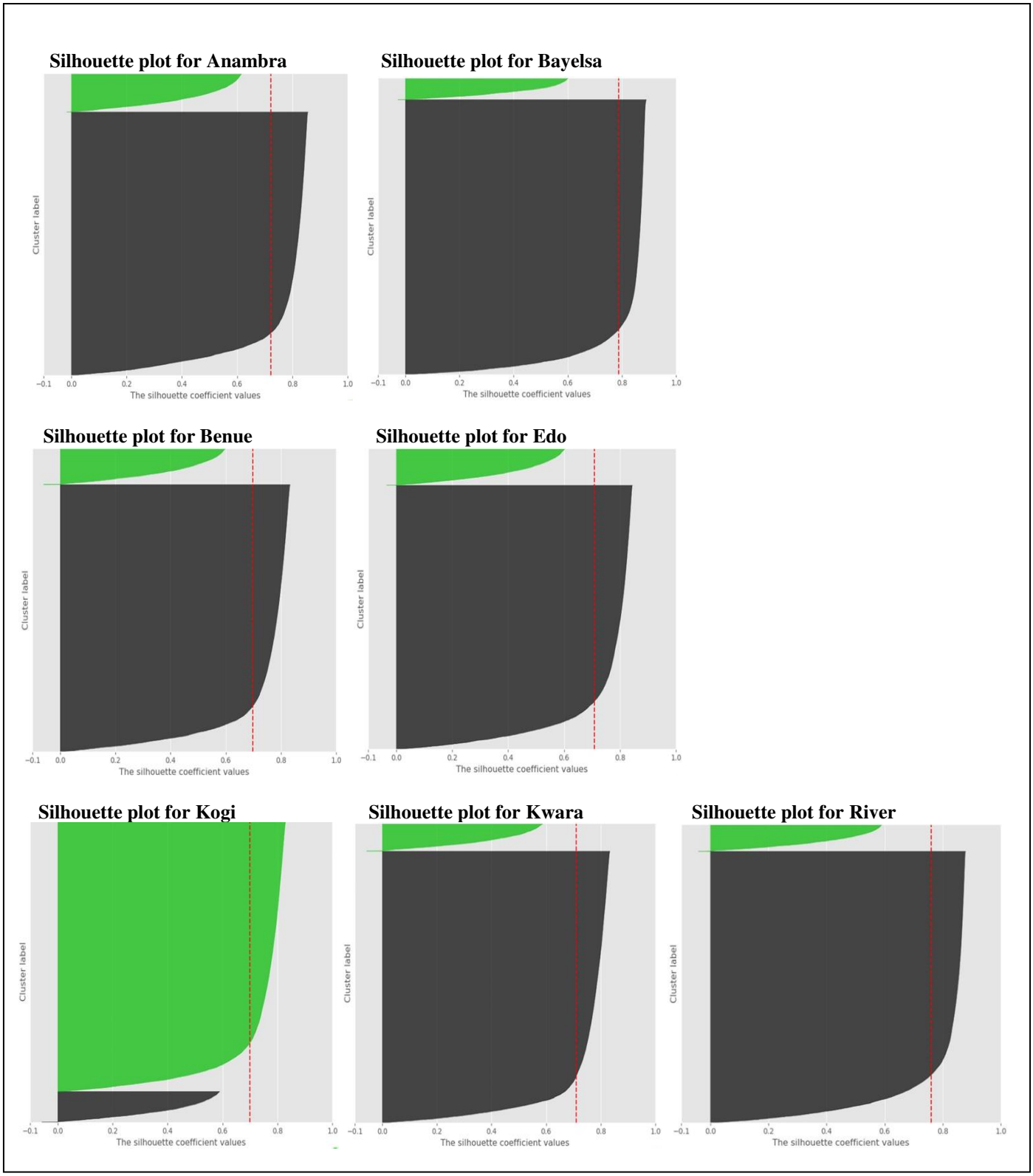

Fig. 7 The Summary of Silhouette Analysis Score with the red line indicating score at k = 2

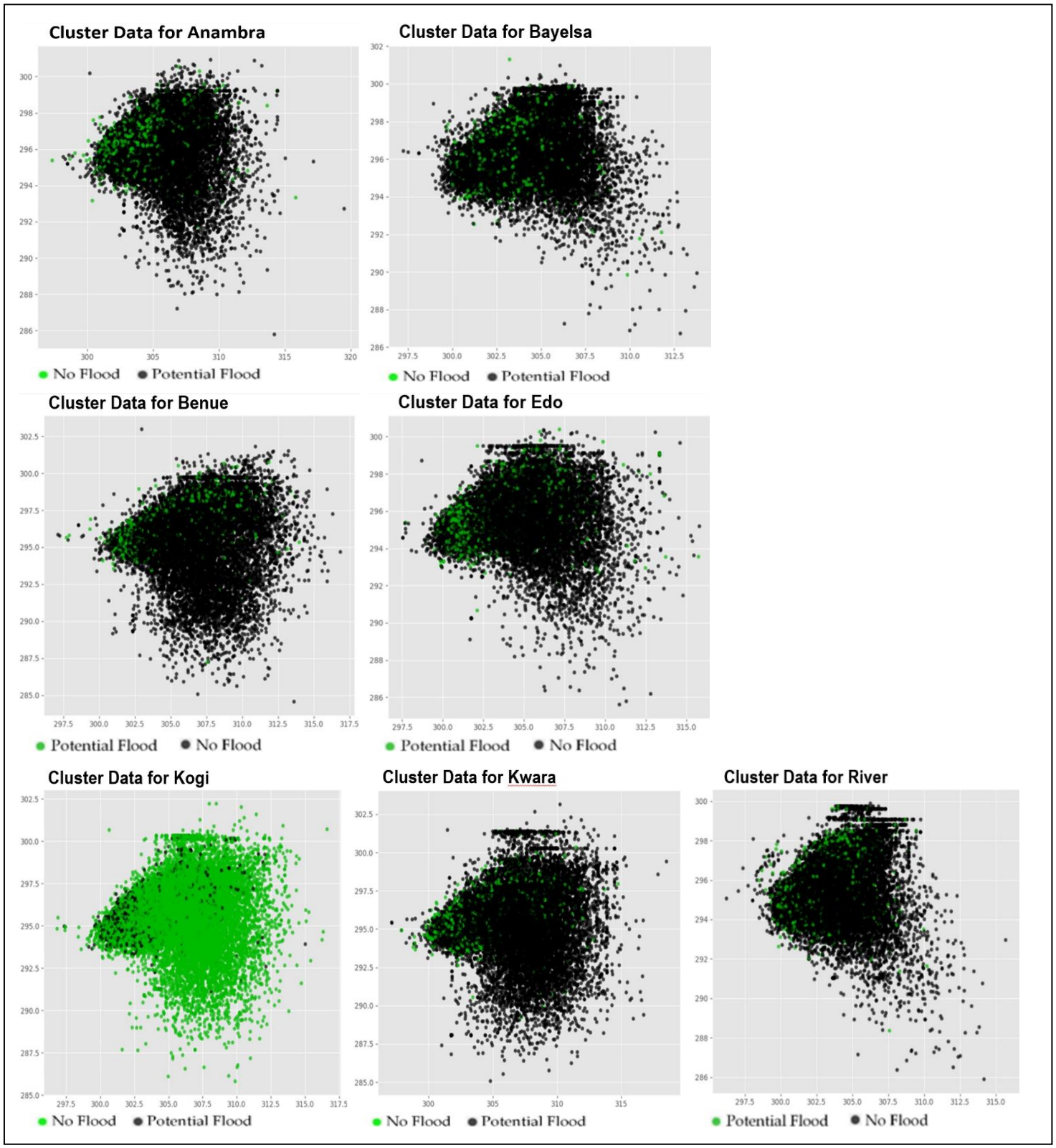

Fig. 8 The summary of clustered feature pattern on each of the seven state

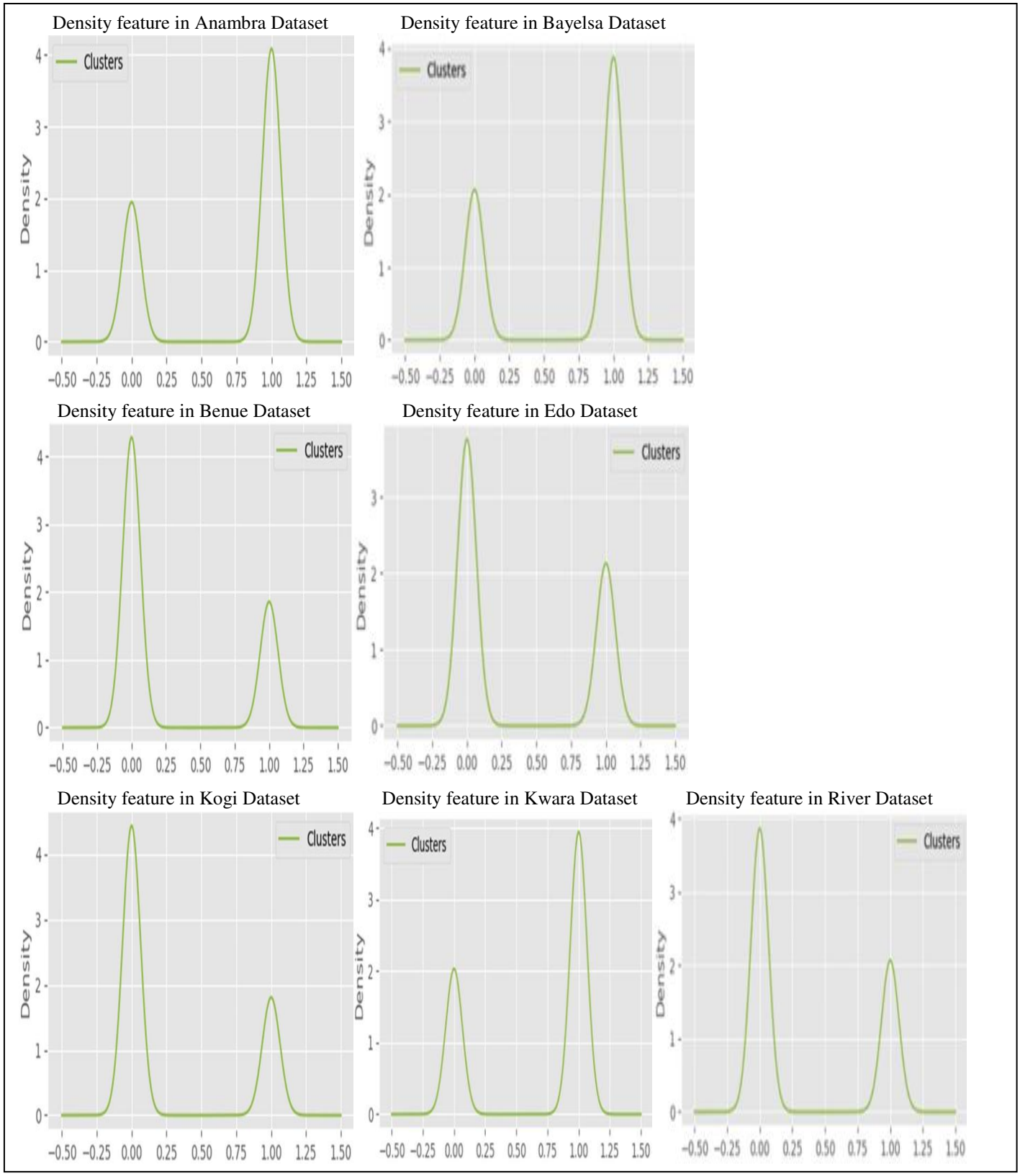

Fig. 9 The summary of the density of features on the four variables parameters in each state

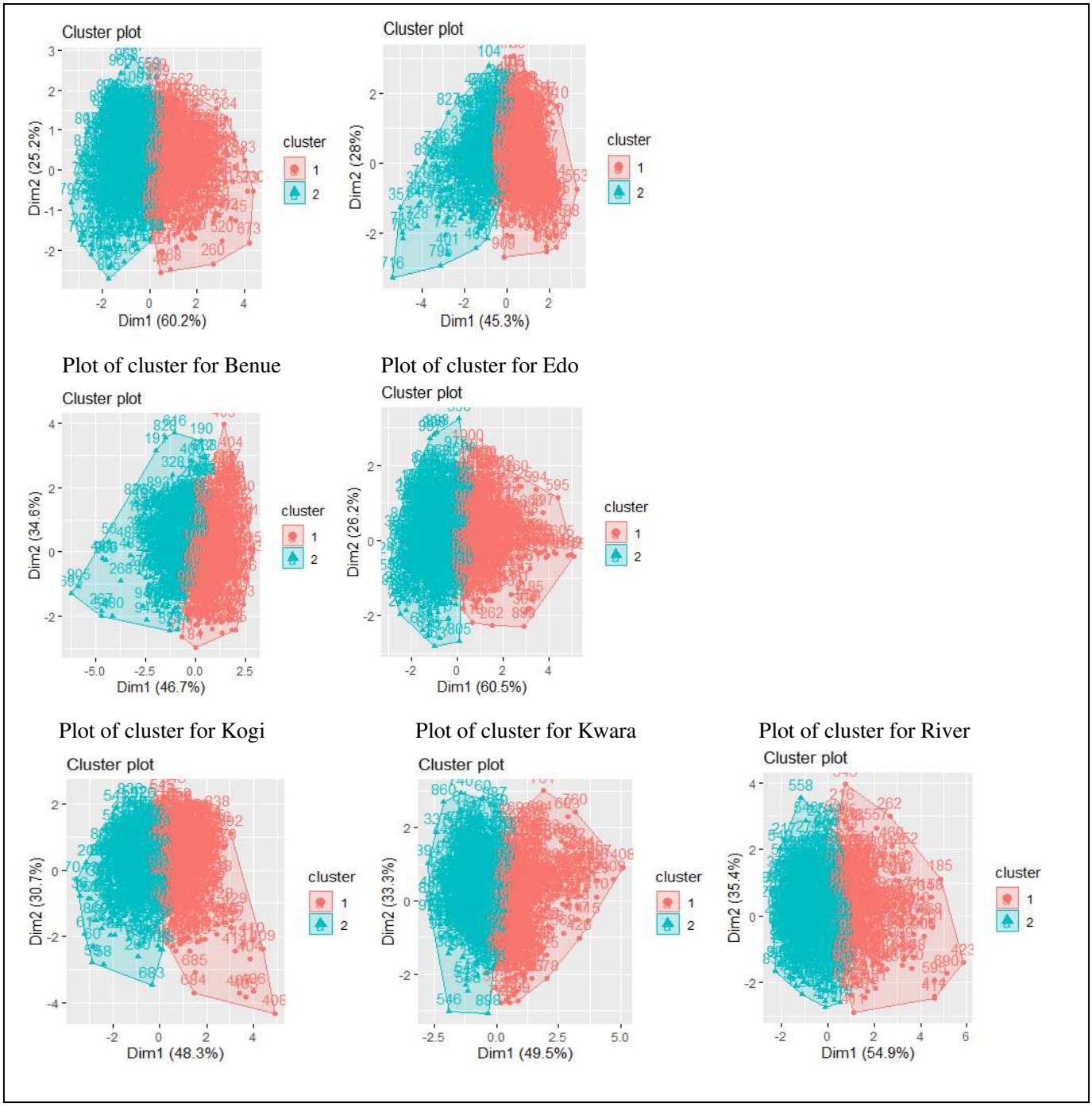

Fig. 10 The summary of the cluster validity feature pattern on each of the seven states

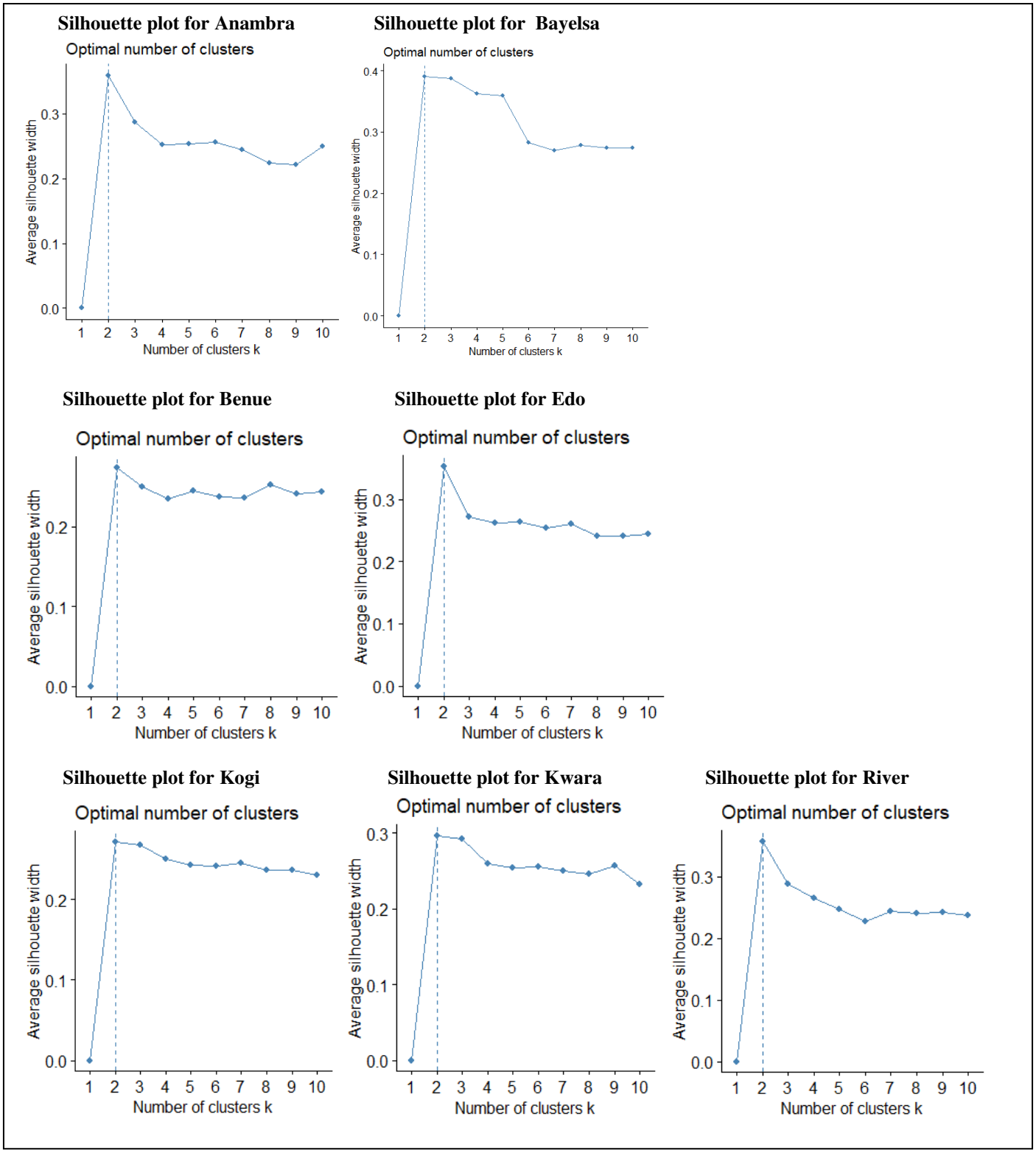

Fig. 11 The visualization of Silhouette's optimal plot

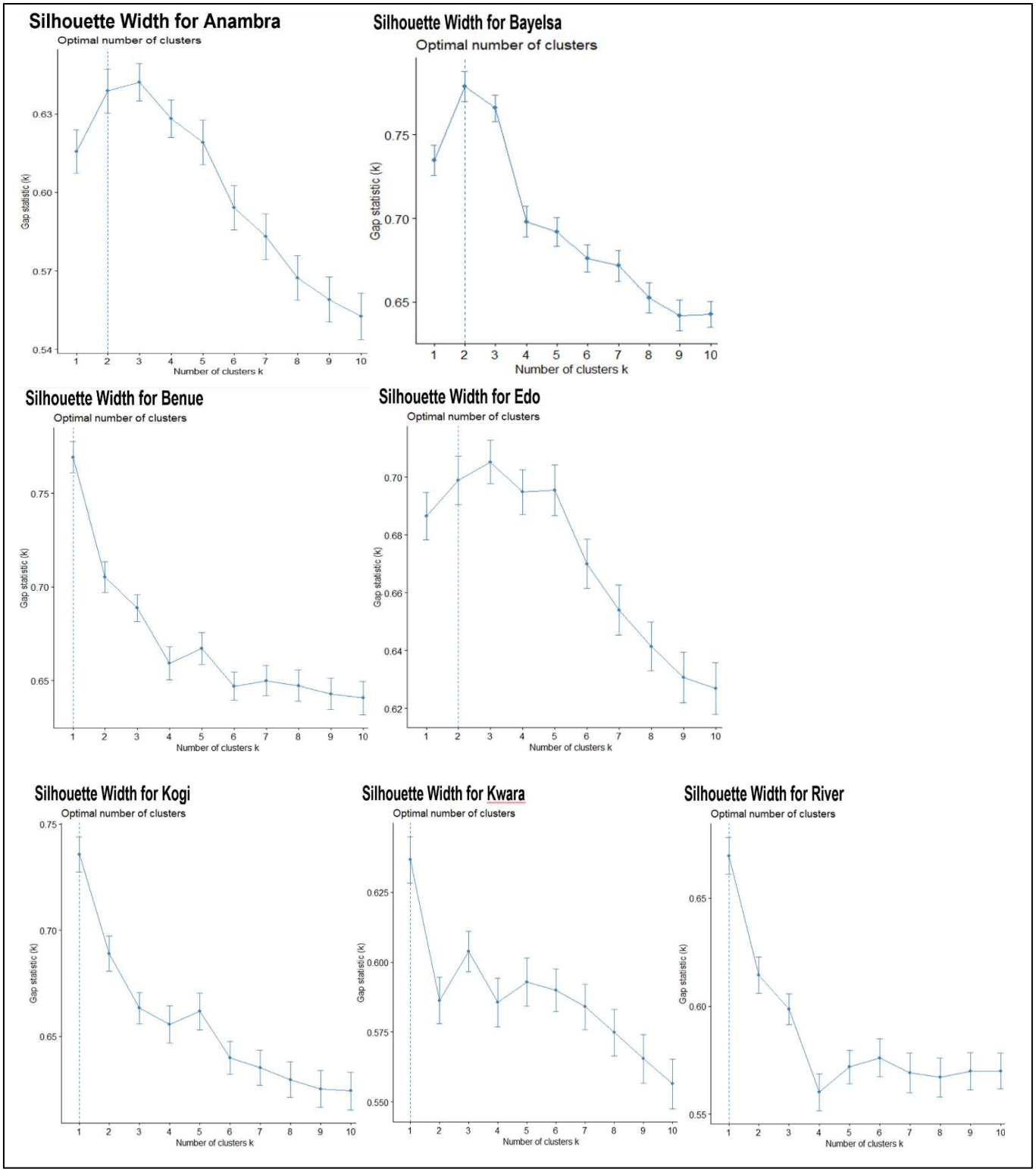

Fig. 12 The visualization of the Gap statistics plot

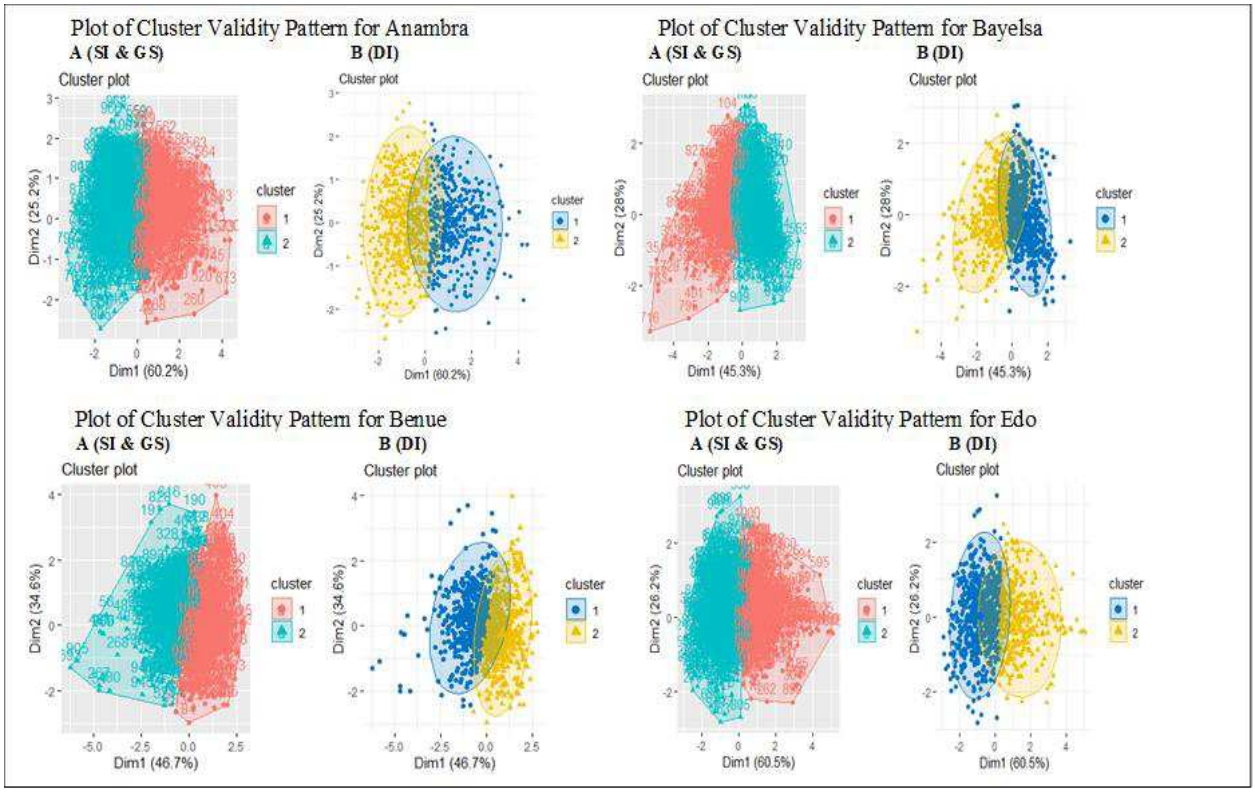

Fig. 13a Summary plots of the cluster validity feature pattern for Anambra, Bayelsa, Benue and Edo

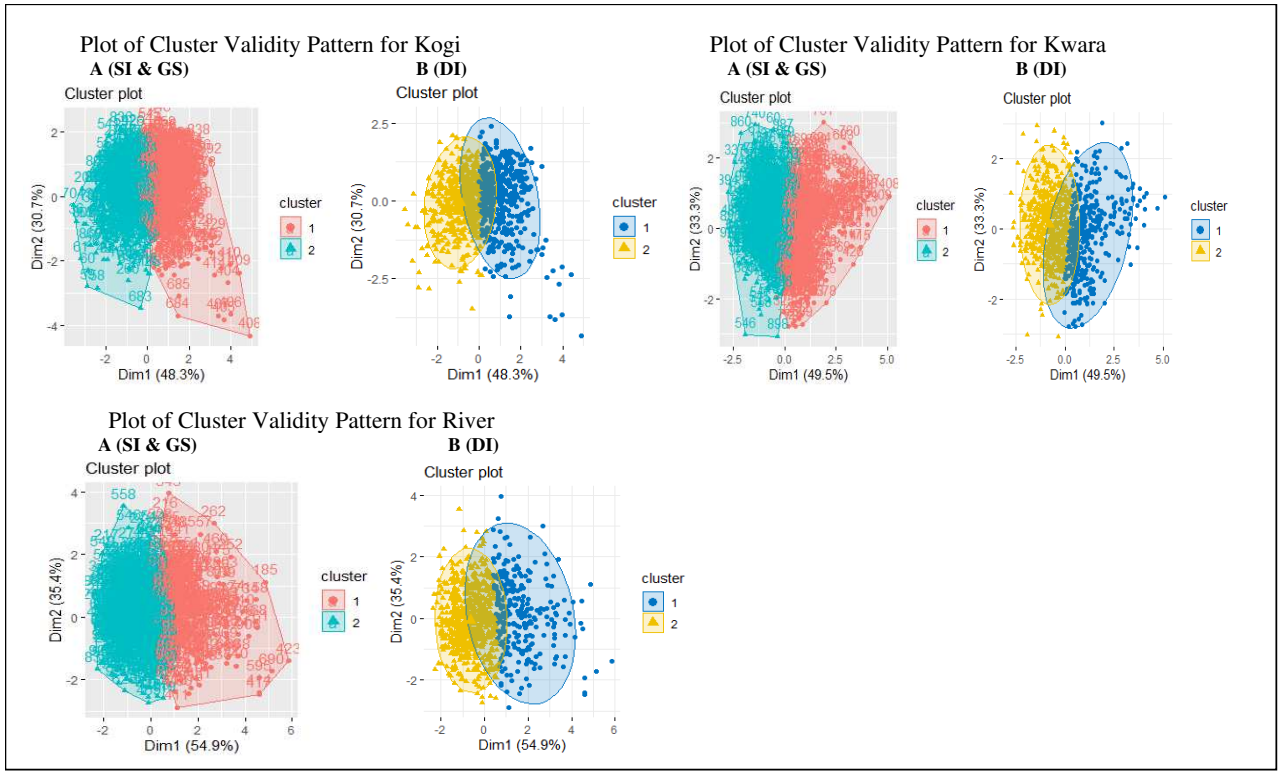

Fig. 13b Summary plots of the cluster validity feature pattern for Kogi, Kwara, and River