University of Ibadan
Journal of Science and Logics in ICT Research

Application of Matrix Multiplication Technique to Google Trends Data for Mining Knowledge on Nigeria's Telecommunications Industry

Adeola O. Opesade

Africa Regional Centre for Information Science, University of Ibadan, Nigeria
ao.opesade@ui.edu.ng, morecrown@gmail.com

Abstract

Data is the raw material of the present Information Age. While there are many sources of big data, the rapid growth of the web and the variety of its data types has made it the largest publicly accessible data source in the world. Google Trends (GT), a web-based data source, has been investigated and analysed by many previous studies. It could however, be observed that these previous studies have mostly analysed GT data based on either time or geographical locations. The present study applies the mathematical principle of matrix multiplication to extend the use of GT data for mining purposes. Dataset derived from the proposed data integration model was evaluated on Nigeria's Global System of Mobile Communication (GSM) operators' (MTN, AIRTEL, GLO, ETISALAT) timeline and geographical search volumes. Supervised learning experiment on the integrated dataset resulting from the proposed model performed very favourably. Further analyses on the resulting dataset showed that search volume for MTN was the most consistent across the 37 Nigerian locations. MTN search also had the highest coverage while AIRTEL had the highest average search volume among the four operators. Analysis of Variance and Post-Hoc tests showed that the search volumes of the four GSM operators were significantly different from one another.

Keywords: *Data integration, Data Science, Global System of Mobile Communication, Google Trends, Matrix Product.*

I. INTRODUCTION

Big data as introduced by Cox and Ellsworth in the late 1990s and cited by Jifa and Lingling [1] describes a phenomenon in which datasets become so large that the memory capacities of data processor are threatened. This description has however, undergone a number of transformations over time. Laney [2] submitted that with E-Commerce, data management challenges have exploded along three dimensions, namely, volume, velocity, and variety. These three constituted the main descriptive features of big data until more recently, when some other ones were discovered such as veracity, variability and complexity. Combining these attributes together, TechAmerica Foundation as cited by Gandomi and Haider [3] defined big data as 'a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of information'.

While there are many sources of big data due to availability of numerous applications and devices that collect information continuously [4], the rapid growth of the web and the variety of its data types have made it the largest publicly accessible data source in the world. Several individuals and

Adeola O. Opesade (2019). Application of Matrix Multiplication Technique to Google Trends Data for Mining Knowledge on Nigeria's Telecommunications Industry, *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),* Vol. 3, No. 1 pp. 1-15
©U IJSLICTR Vol. 3, No. 1 June 2019

organizations have their activities through the www, leaving behind traceable digital footprints. Treatment of these footprints with the appropriate Big Data architecture could lead to discovery of human and organizational behaviour, decisions and intentions, which can assist in examining important socioeconomic changes and trends [5].

The most frequently utilized providers of web data are social media such as Twitter and Facebook and search engines such as Google and Yahoo! [6]. Among these data sources, Google has consistently been put at the top. According to Internet live stats [7], Google was reported to have processed more than 66.7 percent of all the online queries in the world in December 2012. In the Year 2009, Google began the release of its users' search queries through a publicly accessible interface named Google Trends (GT) [8], which offers a search index for the volume of queries based on geographical locations and time, right from the Year 2004 [9]. The index for each query phrase, always a number between 0 and 100, supplies an index which is calculated as the search volume for the query in a given geographical location divided by the total number of queries in that region at a given point in time. The reported numbers therefore, are a demonstration of search interest relative to the highest point on the chart for the given region and time [9, 8].

Many studies have utilised GT data in diverse research areas such as Epidemiology [10, 11, 12, 13], Finance and Economics [6, 8, 9, 14, 15], Webometrics [16] among others. These previous studies reported positively on the usefulness of Google Trends as a veritable source of data for elucidating useful information about human behaviour online. It could however, be observed that these previous studies have mostly analysed GT data based either on time [10, 11, 12, 13] or location [6, 8, 9, 14, 15, 16]. Are there some other forms of useful knowledge that can be mined from Google Trends data?

Data is the raw material of the present information age; the need to mine as much useful information and knowledge from data and to discover trends and patterns from massive stores of big data has led to the emergence of a new field called Data Science [1]. Although there is no consensus on the definition of Data Science yet, it is regarded as a new science, whose research objectives are different from those of other more established branches of science [17]. It employs techniques and theories drawn from many fields within the context of Mathematics, Statistics, Information Science, and Computer Science to understand and analyse actual phenomena with data [18]. According to Jifa and Lingling [1], a Data Scientist can think outside the box, always concerned with what can be made of lots of available data. These suggest that researching into innovative means of handling data in such a way that useful information and knowledge can be mined in a creative and systematic manner, with the use of appropriate tools and techniques is germane in Data Science.

Linear algebra is a field of mathematics that could be called the mathematics of data [19]. In Linear Algebra, matrix multiplication is a binary operation that produces a matrix from two matrices. When two linear maps are represented by matrices, their matrix product represents the composition of the two maps. The purpose of the present study is to extend the use of GT Data beyond the present practice by applying the mathematical principle of matrix multiplication to integrate separately existing timeline and location datasets into a single dataset, useful for temporal-spatial data mining. The specific objectives of the present study are to propose a model for generating temporal-spatial data from Google Trends timeline and geographical data; evaluate the performance of the proposed temporal-spatial datasets through machine learning classification experiments; mine useful information from the proposed temporal-spatial dataset through the use of appropriate supervised machine learning algorithm; and lastly to mine useful information from the proposed temporal-spatial dataset through the use of appropriate unsupervised machine learning algorithm.

II. METHODOLOGY

Problem formulation

Given n terms representing web users' interests and whose time-based and location-based datasets are downloaded from Google Trends. The goal is to integrate the two datasets into a single temporal-spatial data such that useful information which otherwise might not be derivable could be available through the use of appropriate data mining techniques.

Data integration modeling

Let users' search interests for n terms be represented as $T = (t_1, t_2, \ldots\ldots\ t_n)$ … (1)

Google Trends time-based dataset for each term be represented as Matrix $M = \{m_1, m_2, \ldots\ m_p\}$; and Google Trends location-based dataset for each term be represented as

Matrix $L = \{l_1, l_2, \ldots\ldots\ l_q\}$ … (2)
where:

n is the number of terms under investigation, p is the number of time sessions in the time-based dataset and q is the no of locations in the location-based dataset.

If Matrix M (time-based dataset) for each search term is transposed such that it has p columns (no. of time sessions), and Matrix L (location-based dataset) contains q rows (no of locations), then their matrix product (proposed temporal-spatial dataset) is a q by p matrix R represented as:

$R = \{r_1, r_2, \ldots\ldots\ldots\ r_q\}$ … (3)

Where each r_i is a p dimensional vector over the time space. Thus, a location's interests in each of the n terms can be represented as a time vector.

Model Evaluation

To evaluate the performance of the proposed data integration model, Google Trends data was collected and integrated for four GSM operators in Nigeria. An experiment was carried out with five (5) supervised learning algorithms in WEKA to discover regularities in the resulting dataset. The process of data collection, integration and classification experiments are presented in this section.

Data collection, integration and aggregation

Data on the search volume of four GSM operators in Nigeria, namely, MTN, Airtel, Glo mobile and Etisalat were collected from Google Trends. To retrieve search volumes on these terms, each operator's name as popularly called by Nigerians 'MTN', 'AIRTEL', 'GLO', 'ETISALAT' was entered successively as search terms on the explore interface of the Google Trends such that country attribute was set to Nigeria, date was set between January 1, 2016 to December 31, 2018, and Classification attribute was set to Internet and Telecoms. Timeline data and geographical (36 states and 1 Federal Capital Territory) data were downloaded in .csv format for each search term.

Timeline dataset for each of the four GSM operators' search volume (157 by 1 matrix) was transposed to form a 1 by 157 matrix using Microsoft Excel. Location-based datasets for each of the four GSM operators' search volume for Nigeria's thirty-six states and the Federal Capital Territory also formed a 37 by 1 matrix. Matrix product was carried out with Microsoft Excel Matrix Multiplication (MMULT) function for dataset pair for each GSM operator to give a 37 by 157 temporal-spatial dataset. The resulting datasets for all the four GSM operators were then aggregated to form a 148 by 157 matrix.

Dataset Pre-processing for Experimentation

To evaluate the performance of the resulting (real) dataset from the data integration model, control datasets were generated and the performances of the real and control datasets were compared through supervised machine learning technique. To achieve this, the dataset that was derived from the data integration model was pre-processed to produce eight different datasets. The pre-processing was carried out such that firstly, the correct GSM operators' names were appended as the class attribute of the resulting matrix to form the research dataset over the entire period (2016-2018). A control dataset was created by generating random numbers between one and four, assigning a GSM operator's name to a specific random number, and appending the derived GSM operators' names as the class attribute to form the control dataset for the entire period (2016-2018). Three datasets (dataset for the Year 2016, Year 2017 and for the Year 2018) were also derived from each of the first two datasets. This was done to compare the performances of the real and control datasets in each year (2016, 2017, 2018) and over the entire period under investigation (2016-2018). The descriptions of the datasets are presented in Table 1.

Table 1: Description of Resulting Datasets for Experimentation

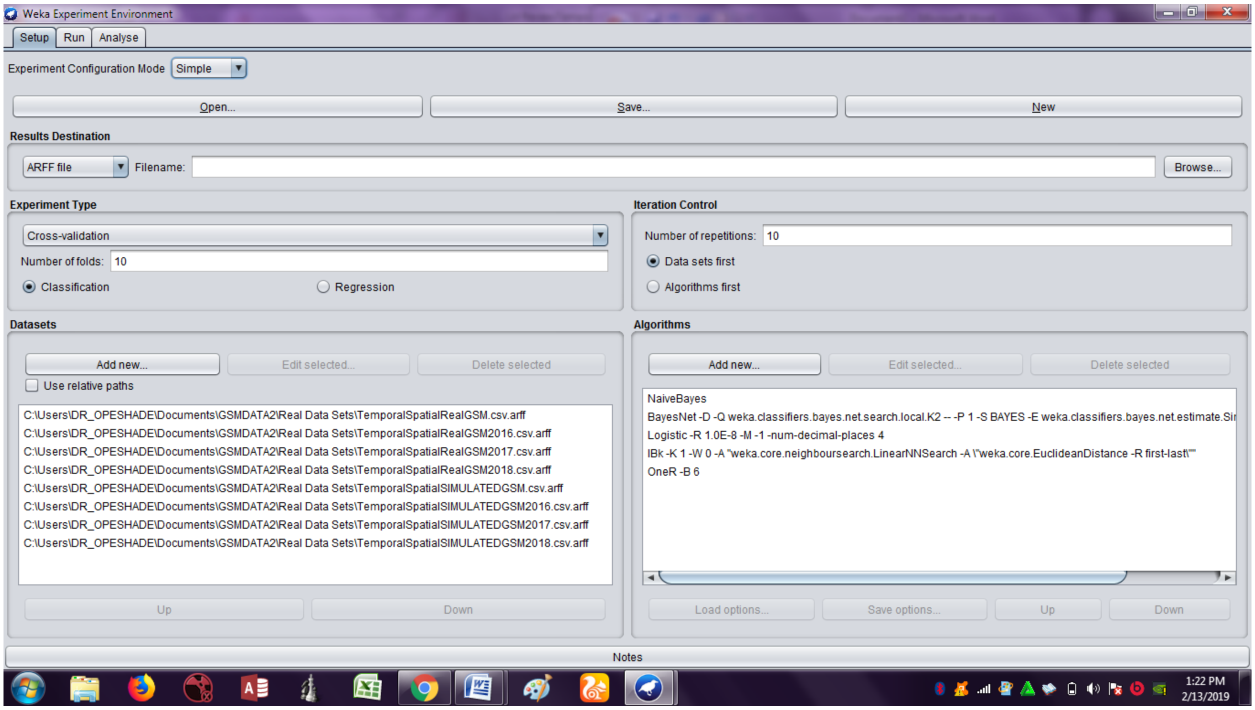
Dataset	**Description**	**Dataset Type**
Dataset 1	• A 148 by 157 matrix • Covers the entire period of research interest **(2016 -2018)** • Instances contain **real class labels**	Real
Dataset 2	• A 52 by 157 matrix • Covers 1 year period of research interest **(2016)** • Instances contain **real class labels**	Real
Dataset 3	• A 53 by 157 matrix • Covers 1 year period of research interest **(2017)** • Instances contain **real class labels**	Real
Dataset 4	• A 52 by 157 matrix • Covers 1 year period of research interest **(2018)** • Instances contain **real class labels**	Real
Dataset 5	• A 148 by 157 matrix • Covers the entire period of research interest **(2016 -2018)** • Instances contain **randomised class labels**	Control
Dataset 6	• A 52 by 157 matrix • Covers 1 year period of research interest **(2016)** • Instances contain **randomised class labels**	Control
Dataset 7	• A 53 by 157 matrix • Covers 1 year period of research interest **(2017)** • Instances contain **randomised class labels**	Control
Dataset 8	• A 52 by 157 matrix • Covers 1 year period of research interest **(2018)** • Instances contain **randomised class labels**	Control

Experimental Setup

The Experimenter interface of the open source Waikato Environment for Knowledge Analysis (WEKA) machine learning tool was used for the supervised learning experimentation. Five machine learning algorithm implementations in WEKA namely NaïveBayes, BayesNet, Logistic, IBK and OneR were used for the purpose of triangulation. The experiment was carried out to determine how well the resulting datasets could be classified into assigned classes of GSM operators.

The assumption is that the higher the performance measures of a real dataset in the classification models, and the greater the variations between a real and control dataset pairs, the higher the evidence that the real datasets represent users' search interests as contained in the timeline and geographical datasets downloaded from GT. If the proposed (real) temporal-spatial datasets perform much better than the control datasets, we conclude that the proposed model can be used for further analyses to mine some useful information contained in the dataset.

The experiment was carried out using ten-fold cross validation repeated ten times. The setting of the experiment is as shown in Fig. 1.



Experiment Configuration Mode: Simple

Results Destination: ARFF file, Filename:

Experiment Type: Cross-validation
Number of folds: 10
● Classification ○ Regression

Iteration Control
Number of repetitions: 10
● Data sets first
○ Algorithms first

Datasets
□ Use relative paths

```
C:\Users\DR_OPESHADE\Documents\GSMDATA2\Real Data Sets\TemporalSpatialRealGSM.csv.arff
C:\Users\DR_OPESHADE\Documents\GSMDATA2\Real Data Sets\TemporalSpatialRealGSM2016.csv.arff
C:\Users\DR_OPESHADE\Documents\GSMDATA2\Real Data Sets\TemporalSpatialRealGSM2017.csv.arff
C:\Users\DR_OPESHADE\Documents\GSMDATA2\Real Data Sets\TemporalSpatialRealGSM2018.csv.arff
C:\Users\DR_OPESHADE\Documents\GSMDATA2\Real Data Sets\TemporalSpatialSIMULATEDGSM.csv.arff
C:\Users\DR_OPESHADE\Documents\GSMDATA2\Real Data Sets\TemporalSpatialSIMULATEDGSM2016.csv.arff
C:\Users\DR_OPESHADE\Documents\GSMDATA2\Real Data Sets\TemporalSpatialSIMULATEDGSM2017.csv.arff
C:\Users\DR_OPESHADE\Documents\GSMDATA2\Real Data Sets\TemporalSpatialSIMULATEDGSM2018.csv.arff
```

Algorithms

```
NaiveBayes
BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.Si
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
OneR -B 6
```

Fig. 1: Experimenter view of WEKA for the datasets and selected algorithms

III. Results and Discussion

Result of Experiment

The models' performances were evaluated at 0.05 (two-tailed) confidence level using six performance measures namely, the true positive rate, false positive rate, Kappa statistic, F-Measure, area under ROC and the Percent correct. The results are as presented in Tables 2a and 2b.

Table 2a: Comparison of TP Rate, FP Rate and Percent Correct measures of the Real (proposed temporal-spatial datasets) and the Control Datasets

Experiment		**TP Rate**		**FP Rate**		**Percent Correct**	
Algorithm	**Dataset**	**Real**	**Control**	**Real**	**Control**	**Real**	**Control**
Naïve Bayes	2016	0.84	0.35	0.10	0.20	78.33	30.80
	2017	0.84	0.25	0.03	0.15	83.40	30.67
	2018	0.88	0.06	0.01	0.05	82.47	28.59
	2016-2018	0.87	0.23	0.02	0.14	83.80	31.04
Bayes Net	2016	0.84	0.28	0.04	0.28	82.93	25.63
	2017	0.84	0.28	0.04	0.28	83.13	25.63
	2018	0.92	0.28	0.01	0.28	83.22	25.63
	2016-2018	0.88	0.28	0.03	0.28	83.80	25.63
Logistic	2016	0.97	0.37	0.00	0.31	94.94	30.12
	2017	0.97	0.37	0.00	0.31	95.08	30.12
	2018	0.97	0.37	0.00	0.31	94.93	30.12
	2016-2018	0.97	0.37	0.00	0.31	94.80	30.12
IBK	2016	0.95	0.43	0.01	0.30	92.17	27.86
	2017	0.97	0.38	0.00	0.30	93.92	25.89
	2018	0.95	0.41	0.00	0.30	93.45	26.65
	2016-2018	0.97	0.41	0.00	0.30	93.99	26.97
OneR	2016	0.82	0.48	0.07	0.30	72.12	28.03
	2017	0.88	0.44	0.04	0.35	79.30	24.87
	2018	0.86	0.44	0.03	0.37	83.68	26.15
	2016-2018	0.86	0.44	0.03	0.32	83.68	26.00

As shown in Table 2a, using Naive Bayes algorithm, the True Positive (TP) rate varies between 0.88 and 0.84 for the proposed temporal-spatial dataset and between 0.35 and 0.06 for the control dataset. The False Positive varies between 0.10 and 0.01 for the proposed temporal-spatial dataset and between 0.20 and 0.05 for the control dataset. While the Percent Correct varies between 83.80 and 78.33 for the proposed temporal-spatial dataset and between 31.04 and 28.59 for the control dataset. For all these three measures, the proposed datasets outperformed the control datasets irrespective of timeframe of the datasets. This trend is consistent across the five algorithms compared in this experiment. It could also be observed that the result of Logistic was the most precise, particularly for the proposed temporal-spatial dataset.

Table 2b: Comparison of Kappa Statistics, F-Measure, ROC Area measures of the proposed temporal-spatial (Real) Datasets and the Control Datasets

Experiment		**Kappa Statistics**		**F- Measure**		**ROC Area**	
Algorithm	**Dataset**	**Real**	**Control**	**Real**	**Control**	**Real**	**Control**
NaïveBayes	2016	0.71	0.09	0.78	0.37	0.93	0.58
	2017	0.78	0.09	0.86	0.30	0.94	0.56
	2018	0.77	0.07	0.91	0.23	0.96	0.53
	2016-2018	0.78	0.09	0.88	0.30	0.95	0.57
BayesNet	2016	0.77	0.00	0.85	0.41	0.94	0.50
	2017	0.77	0.00	0.85	0.41	0.94	0.50
	2018	0.78	0.00	0.93	0.41	0.97	0.50
	2016-2018	0.78	0.00	0.88	0.41	0.96	0.50
Logistic	2016	0.93	0.06	0.99	0.33	1.00	0.59
	2017	0.93	0.06	0.99	0.33	1.00	0.59
	2018	0.93	0.06	0.99	0.29	1.00	0.59
	2016-2018	0.93	0.06	0.99	0.32	0.99	0.59
IBK	2016	0.90	0.04	0.96	0.38	0.96	0.56
	2017	0.92	0.01	0.98	0.34	0.97	0.55
	2018	0.91	0.02	0.97	0.35	0.95	0.55
	2016-2018	0.92	0.03	0.98	0.36	0.97	0.54
OneR	2016	0.69	0.04	0.80	0.41	0.87	0.59
	2017	0.72	0.01	0.88	0.36	0.92	0.54
	2018	0.78	0.02	0.88	0.34	0.92	0.54
	2016-2018	0.78	0.01	0.88	0.37	0.92	0.56

Also as shown in Table 2b, for the three measures considered, that is, kappa statistics, F-measure and ROC Area, the proposed temporal-spatial (real) datasets outperformed the control datasets irrespective of timeframe of the datasets. The results of Logistic were also the most precise particularly for the proposed temporal-spatial dataset. Using Naive Bayes algorithm, Kappa Statistics vary between .78 and .71 for the proposed temporal-spatial dataset and between 0.09 and 0.07 for the control dataset. The F-measure varies between 0.91 and 0.78 for the proposed temporal-spatial dataset and between 0.37 and 0.23 for the control dataset. While the area under ROC varies between .96 and .93 for the proposed temporal-spatial dataset and between 0.58 and 0.53 for the control dataset. This trend of the proposed datasets outperforming the control datasets is also consistent across the five algorithms compared in this experiment.

Mining the Resulting Temporal-Spatial Dataset

Having evaluated and found favourable the performance of the resulting dataset from the proposed data integration model, attempts were made to mine the dataset for useful information. Findings from the supervised and unsupervised machine learning techniques are presented in this section.

Mining temporal-spatial dataset with logistic (classification) algorithm

Logistic algorithm was applied to Dataset 1 to mine some relevant information from the dataset.

Based on 10-fold cross validation repeated ten times, the model's accuracy and kappa statistic were 93.9% and 0.92 respectively. The detailed performance of the Function Logistic model that classified temporal-spatial dataset of Nigerian GSM operator's search into their respective classes is as shown in Table 3a while the confusion matrix is shown in Table 3b.

Table 3a: Detailed Performance of the Classifier

TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**	**Class**
0.973	0.000	1.000	0.973	0.986	0.982	0.993	0.988	MTN
0.946	0.000	1.000	0.946	0.972	0.964	0.995	0.987	AIRTEL
0.892	0.018	0.943	0.892	0.917	0.891	0.992	0.979	GLO
0.946	0.063	0.833	0.946	0.886	0.848	0.991	0.974	ETISALAT
0.939	0.020	0.944	0.939	0.940	0.921	0.993	0.982	Weighted Avg.

Table 3b: Confusion Matrix of Algorithm's Classification

MTN	**AIRTEL**	**GLO**	**ETISALAT**	**Classified as**	**Accuracy (%)**
36	0	0	1	MTN	97.3
0	35	0	2	AIRTEL	94.6
0	0	33	4	GLO	89.2
0	0	2	35	ETISALAT	94.6

The accuracy of classification are 97.3 for MTN, 94.6 for AIRTEL, 89.2 for GLO and 94.6 for ETISALAT. This shows that across the 36 states and FCT of Nigeria, the greatest consistency exists in the search volume for MTN, followed by AIRTEL and ETISALAT, and then lastly GLO. It could also be observed that GLO search volume for 4 Nigerian locations were misclassified as ETISALAT and two of ETISALAT are classified as being GLO. This depicts a relatively greater level of similarity in the search volumes of GLO and ETISALAT. Furthermore, it could be observed that false drops from MTN, AIRTEL and GLO are on ETISALAT. This depicts that search volume in those misclassified instances appear as those of ETISALAT.

Mining temporal-spatial dataset with k-means (clustering) algorithm

The resulting (real) dataset was decomposed into the four GSM classes and the class labels were removed from each GSM operator's dataset. Unsupervised K-Means algorithm was used to cluster the geographical locations contained in the dataset of each GSM operator into their inherent classes based on their search volumes over time. The purpose of the clustering was to mine useful information on the popularity (search volume and search coverage) of these GSM operators across Nigeria's 36 states and FCT. Setting the K parameter of the algorithm to 4, the distribution of the centroids of the four clusters of each GSM operator and the descriptive statistics (average, maximum and minimum values) of each cluster are presented in Figs. 2 - 5 and Tables 4 - 7.

i. Popularity of MTN in terms of search volume and coverage

Distribution of the centroids of the four clusters of MTN temporal-spatial dataset is as shown in Fig. 2 while the descriptive statistics of these clusters are presented in Table 4.

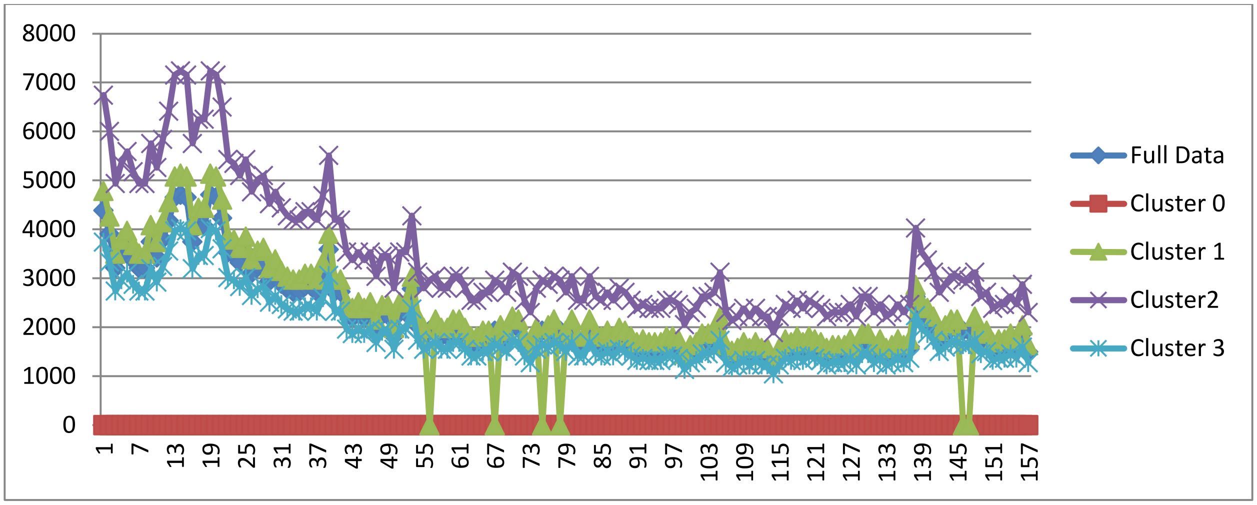


Fig. 2: Distribution of MTN Clusters' Centroids

Table 4: Descriptive statistics of MTN Clusters

	Full Data	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Clustered Instances	37 (100%)	1 (2.7%)	15 (40.5%)	4 (10.8%)	17 (45.9%)
Average	2226.195	0	2441.197	3419.924	1897.987257
Max	4711.568	0	5139.2	7238	4016.9412
Min	1231.432	0	1343.2	1891.75	1049.8824

As derived from Fig. 2 and Table 4, Cluster 2 having four locations has the highest search volume as depicted by its average, maximum and minimum values which are 3419.924, 7238, 1819.75 respectively. This is followed by Cluster 1(15 locations) and then, Cluster 3 (17 locations) and lastly Cluster 0 having only one location with average, maximum and minimum values which are 0, 0, 0 respectively. This shows that MTN has four locations where its search volume is the highest and one location with the least search volume while majority of the locations (32) are in between the two extremes.

ii. Popularity of AIRTEL in terms of search volume and coverage

Distribution of the centroids of the four clusters of AIRTEL temporal-spatial dataset is as shown in Fig. 3 while the descriptive statistics of these clusters are presented in Table 5.

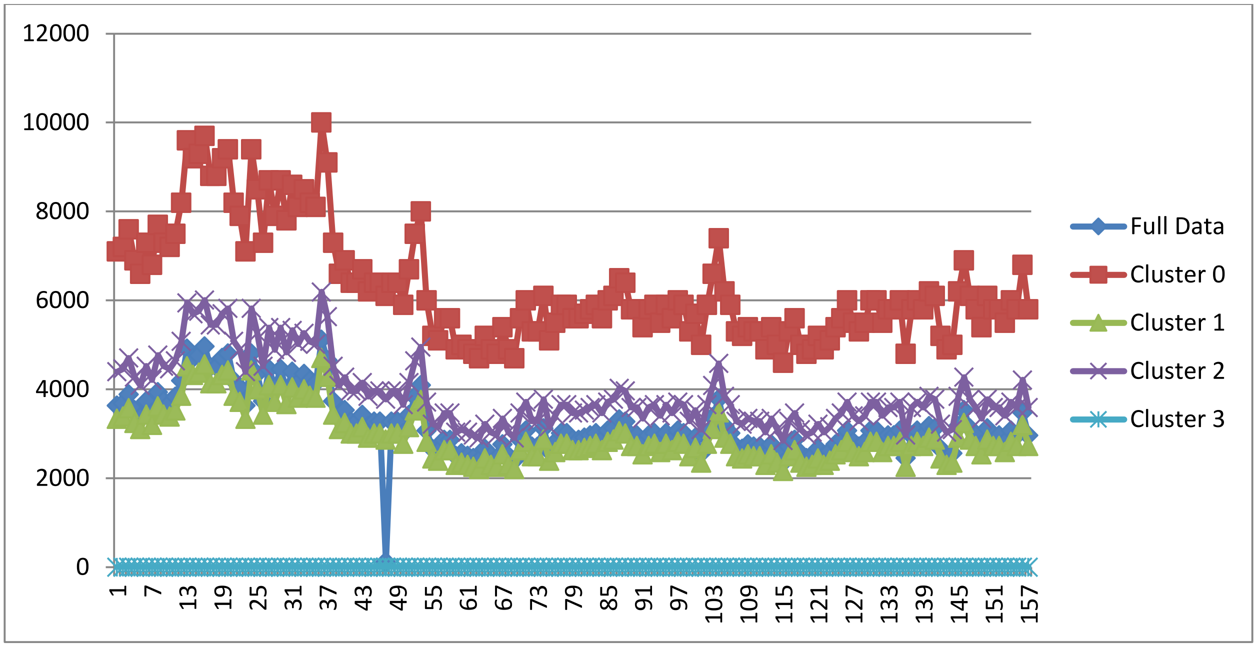


Fig. 3: Distribution of AIRTEL Clusters' Centroids

Table 5: Descriptive statistics of AIRTEL Clusters

	Full Data	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Clustered Instances	37 (100%)	1 (3%)	21 (57%)	13 (35%)	2 (5%)
Average	3208.344	6308.28	2967.896	3906.281	0
Max	5116.216	10000	4704.762	6192.308	0
Min	120.8919	4600	2164.191	2848.462	0

As derived from Fig. 3 and Table 5, Cluster 0 having only one location has the highest search volume as depicted by its average, maximum and minimum values which are 6308.28, 10,000, and 4600 respectively. This is followed by Cluster 2 (13 locations) and then, by Cluster 1 (21 locations) and lastly Cluster 3 having two locations with average, maximum and minimum values which are 0, 0 and 0 respectively. This shows that AIRTEL has only one location where its search volume is the highest. Two locations have the least search volume while majority of the locations (34) are in between the two extremes. It could also be observed that although the number of locations with the highest search volumes in MTN is higher than that of AIRTEL being 4 and 1 respectively, the average search volume of AIRTEL is higher than that of MTN.

iii. Popularity of ETISALAT in terms of search volume and coverage

Distribution of the centroids of the four clusters of ETISALAT temporal-spatial dataset is as shown in Fig. 4 while the descriptive statistics of these clusters are presented in Table 6.

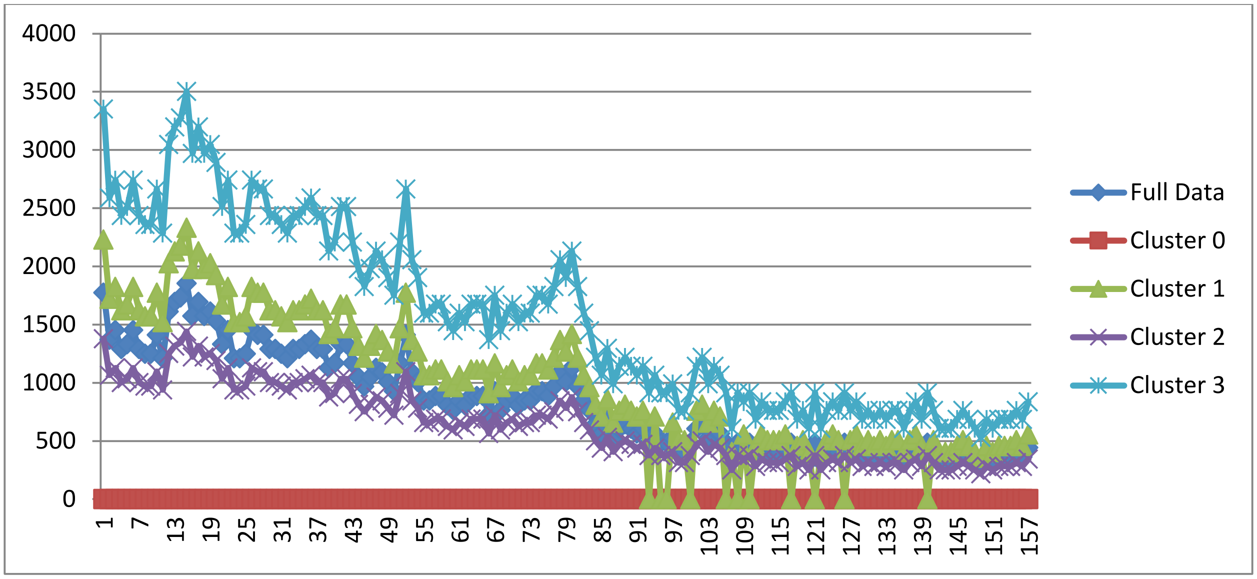


Fig. 4: Distribution of ETISALAT Clusters' Centroids

Table 6: Descriptive statistics of ETISALAT Clusters

	Full Data	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Clustered Instances	37 (100%)	6 (16.2%)	13 (35.1%)	12 (32.4%)	6 (16.2%)
Average	826.7771	0	1071.83	642.4331	1561.659
Max	1854.919	0	2331.846	1441.333	3503.667
Min	282.2703	0	354.8462	219.3333	533.1667

As derived from Fig. 4 and Table 6, Cluster 3 having six locations has the highest search volume as depicted by average, maximum and minimum values which are 1561.659, 3503.667, 533.1667 respectively. This is followed by Cluster 1 (13 locations), Cluster 2 (12 locations) and lastly Cluster 0 having six locations with average, maximum and minimum values which are 0, 0, 0 respectively. The remaining twenty five are distributed in between the two extremes. This shows that ETISALAT's search volumes are more evenly distributed (among the four clusters) than those of MTN and AIRTEL. It could also be observed that the cluster with the highest search volume in ETISALAT has less values when compared with MTN's cluster with the highest volume.

iv. Popularity of GLO in terms of search volume and coverage

Distribution of the centroids of the four clusters of GLO temporal-spatial dataset is as shown in Fig. 5 while the descriptive statistics of these clusters are presented in Table 7.

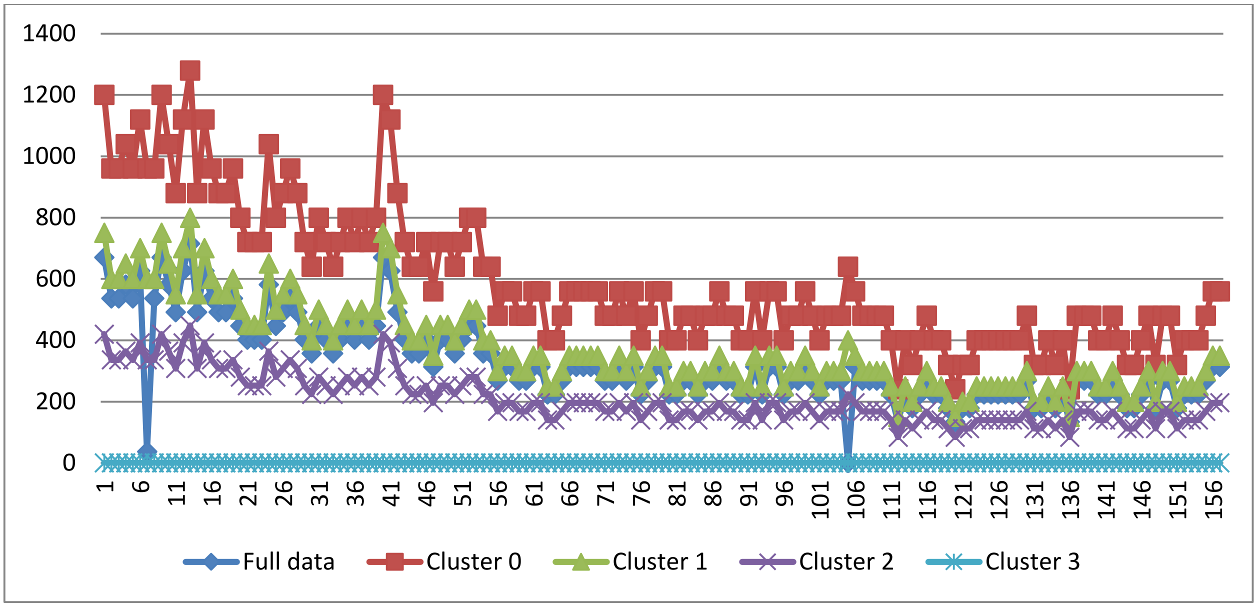


Fig. 5: Distribution of GLO Clusters' Centroids

Table 7: Descriptive statistics of GLO Clusters

	Full Data	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Clustered Instances	37 (100%)	6 (16%)	19 (51%)	8 (22%)	4 (11%)
Average	327.1937	591.5924	369.7452	207.0573	0
Max	715.2432	1280	800	448	0
Min	36.4324	240	150	84	0

As derived from Fig. 5 and Table 7, Cluster 0 having six locations has the highest search volume as depicted by average, maximum and minimum values which are 591.5924, 1280, 240 respectively. This is followed by Cluster 1 (19 locations), Cluster 2 (8 locations) and lastly Cluster 3 having four locations with average, maximum and minimum values which are 0, 0, 0 respectively. The remaining twenty seven are distributed between the two extremes. This shows that GLO search volumes are also more evenly distributed than those of MTN and AIRTEL with six and four locations having the highest and the least search volumes respectively. It could also be observed that the cluster with the highest search volume in GLO has less values when compared with ETISALAT's highest volume cluster.

Statistical Comparison of GSM Operators' Search Volumes

To determine the relative performances of the GSM operators search volumes over the period of study, an inferential statistical test was carried out to determine if there were significant differences in the average search volumes (based on cluster labels) of the four GSM operators from the Nigerian 36 states and FCT. Analysis of Variance (ANOVA) test was carried out using Statistical Package for Social Sciences (SPSS) software to test the hypothesis stated below.

Hypothesis: There are no significant differences in the search volumes of four GSM operators (MTN, AIRTEL, ETISALAT, GLO) in Nigeria. Result of the Test of Hypothesis is as presented in Table 8a.

Findings on the test of hypothesis

Table 8a: Result of the Test of Hypothesis

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.934E8	3	6.446E7	153.590	.000
Within Groups	6.043E7	144	419663.822		
Total	2.538E8	147			

As shown in Table 8a, the result of the Analysis of Variance using ANOVA shows that F=153.590, df=3 Sig. = 0.000. Since the p-value <0.05, the null hypothesis is rejected. There is therefore, a significant difference in the search volumes of at least a pair of the operators. To determine the relative differences among the four operators, Tukey HSD Post-Hoc test was carried out, the result is presented in Table 8b.

Table 8a: Tukey HSD Result

CLASS	N	Subset for alpha = 0.05			
		1	2	3	4
GLO	37	330.5722			
ETISALAT	37		838.1876		
MTN	37			2231.4442	
AIRTEL	37				3227.4527

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 37.000.

The output of the Post-Hoc test shows that the search volumes are different one from another, with AIRTEL having the highest mean (3227.4527) followed by MTN (2231.4442), ETISALAT (838.1876) and lastly GLO (330.5722).

IV. Conclusion

The present study applied the principle of matrix multiplication to model and generate temporal-spatial data from Google Trend's separately provided time-based and location-based data. Supervised machine learning experiment carried out to investigate the performance of the generated dataset shows that the proposed temporal-spatial dataset performed very favourably. MTN has the largest coverage (36 states) followed by AIRTEL (35 states), GLO (32 states) and lastly ETISALAT (30 states). Search volume for MTN is the most consistent across the 37 Nigerian locations followed by AIRTEL and ETISALAT and lastly GLO. Further analysis showed that AIRTEL has the highest search volume followed by MTN, ETISALAT and lastly GLO. Analysis of Variance and Post-Hoc tests showed that the search volumes of these four GSM operators were significantly different from one another.

This study has contributed to knowledge by extending the possibilities of the use of Google trends data in research and web analytics. It has also provided useful information on Nigerians' relative search interests in four GSM operators in the country.

REFERENCES

[1] Jifa G. and Lingling Z. (2014). Data, DIKW, Big data and Data science. 2nd International Conference on Information Technology and Quantitative Management, ITQM 2014, Procedia Computer Science 31. p. 814 – 821.

[2] Laney D. (2001). 3D Data Management: Controlling data volume, velocity, and variety. Application Delivery Strategies. [cited 2019 Feb 14]. Available from http://blogs.gartner.com/douglaney/files/2012/01/ad949-3D-Data-Management Controlling-Data-Volume-Velocity-and-Variety.pdf.

[3] Gandomi A. and Haider M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management. 35: 137–144.

[4] Bhadani A and Jothimani D. (2016). Big data: Challenges, opportunities and realities. In: Singh MK, Kumar DG, editors. Effective big data management and opportunities for implementation Pennsylvania: IGI Global; p. 1-24. Available from https://arxiv.org/pdf/1705.04928.

[5] Blazquez D. and Domenech J. (2017). Big Data sources and methods for social and economic analyses. Technological Forecasting & Social Change. Available from http://dx.doi.org/10.1016/j.techfore.2017.07.027.

[6] Kristoufeka, L. (2013). Can Google Trends search queries contribute to risk diversification? Scientific Reports. 3(2713), DOI:10.1038/srep02713. Available from http://www.nature.com/srep/2013/130919/srep02713/full/srep02713.html

[7] Internet live stats (2014). Google search statistics [Internet]. Available from http://www.internetlivestats.com/google-search-statistics/

[8] Carrière-Swallow Y. and Labbé F. (2013). Nowcasting with Google Trends in an emerging market. Journal of Forecasting. 32(4): 289-298.

[9] Wu L. and Brynjolfsson E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales in economic analysis of the digital economy. In: Goldfarb A, Greenstein SM, Tucker CE, editors. Economic analysis of the digital economy. p.89 – 118. University of Chicago Press. Available from http://www.nber.org/chapters/c12994

[10] Carneiro H. A. and Mylonakis E. (2009). Google Trends: A web-based tool for real-time surveillance of disease outbreaks. Clinical Infectious Disease. 49 (15): 57–64.

[11] Seifter A., Schwarzwalder A., Geis K. and Aucott J. (2010). The utility of “Google Trends” for epidemiological research: Lyme disease as an example. Geospatial Health. 4(2): 135-137.

[12] Malik M. T., Gumel A., Thompson L. H., Strome T. and Mahmud S. M. (2009). “Google Flu Trends” and Emergency Department triage data predicted the 2009 pandemic H1N1 waves in Manitoba. Canadian Journal of Public Health. 2011; 102(4): 294-297.

[13] Kang M., Zhong H,, He J., Rutherford S. and Yang F. (2013). Using Google Trends for influenza surveillance in South China. PLoS ONE. 2013; 8(1): e55205. doi:10.1371/journal.pone.0055205

[14] Vosen S. and Schmidt T. (2011). Forecasting private consumption: Survey-based indicators vs. Google Trends. Journal of Forecasting. 2011; 30(6): 565-578.

[15] Choi H. and Varian H. (2012). Predicting the present with Google Trends. Economic Records. 2012; 88(1): 2-9. https://doi.org/10.1111/j.1475-4932.2012.00809.x

[16] Vaughan L. and Romero-Frías E. (2014). Web search volume as a predictor of academic fame: An exploration of Google Trends. Journal of the American Society For Information Science and Technology. 65(4): 707-720

[17] Zhu Y. and Xiong Y. (2015). Defining Data Science Beyond the study of the rules of the natural world as reflected by data [Internet]. Available from https://arxiv.org/ftp/arxiv/papers/1501/1501.05039.pdf.

[18] Chikio H. (1998). "What is Data Science? Fundamental Concepts and a Heuristic Example". In: Hayashi C, Yajima K, Bock H, Ohsumi N, Tanaka Y, Baba Y. Studies in Classification, Data Analysis, and Knowledge Organization. Japan: Springer. p. 40–51. doi:10.1007/978-4-431-65950-1_3 ISBN 9784431702085.

[19] Brownlee J. (2018). 5 Reasons to Learn Linear Algebra for Machine Learning [Internet]. 2018. Available from https://machinelearningmastery.com/why-learn-linear-algebra-for-machine-learning/