World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:13, No:2, 2019

Localization of Geospatial Events and Hoax Prediction in the UFO Database

Harish Krishnamurthy, Anna Lafontant, Ren Yi

***Abstract*—**Unidentified Flying Objects (UFOs) have been an interesting topic for most enthusiasts and hence people all over the United States report such findings online at the National UFO Report Center (NUFORC). Some of these reports are a hoax and among those that seem legitimate, our task is not to establish that these events confirm that they indeed are events related to flying objects from aliens in outer space. Rather, we intend to identify if the report was a hoax as was identified by the UFO database team with their existing curation criterion. However, the database provides a wealth of information that can be exploited to provide various analyses and insights such as social reporting, identifying real-time spatial events and much more. We perform analysis to localize these time-series geospatial events and correlate with known real-time events. This paper does not confirm any legitimacy of alien activity, but rather attempts to gather information from likely legitimate reports of UFOs by studying the online reports. These events happen in geospatial clusters and also are time-based. We look at cluster density and data visualization to search the space of various cluster realizations to decide best probable clusters that provide us information about the proximity of such activity. A random forest classifier is also presented that is used to identify true events and hoax events, using the best possible features available such as region, week, time-period and duration. Lastly, we show the performance of the scheme on various days and correlate with real-time events where one of the UFO reports strongly correlates to a missile test conducted in the United States.

***Keywords*—**Time-series clustering, feature extraction, hoax prediction, geospatial events.

I. Introduction

THE data used in this research are collected and made public by the National UFO Reporting Center launched in 1974. The NUFORC site hosts an extensive database of UFO sighting reports that are submitted either online or through a 24-hour telephone hotline. The data undergo an internal quality check by NUFORC staff before being made public and, at the moment, present one of the most comprehensive UFO reports databases available online. It provides the following information: Date/Time, City, State, Shape, Duration, Summary, and Posting date. The data get occasionally used for local news reports as well as a broader-level reporting. Due to its accessibility online, the dataset has been used for various forms of visualization and analysis. Notably, there has been a heavy focus on mapping and visualization of the report sightings. One of the key challenges of working with the UFO reports data is the data quality and credibility of the source, and in particular, the necessary distinction between a UFO and an IFO (identifiable flying object); "Studies of UFO data routinely include reports of meteors, fireballs, and other conventional object" [5]. The inevitable presence of IFO reports in the dataset can, in fact, be considered as an added value, since the non-UFO reports are still indicative of actual events taking place. Therefore, this analysis focuses on events that are reported as UFOs, regardless of them being an alien activity or in future recognized as an IFO. In addition to general reporting trends, the analysis of NUFORC data can offer insight into UFO perception and their validity, as some of the latter are labeled to be hoax reports by NUFORC.

Lastly, we exhibit an event localization method that uses unsupervised clustering and time-series analysis methods to identify and potentially track events that trigger the UFO reports. The existing approaches to time series clustering usually consist of temporal-proximity-based clustering , representation-based clustering and model-based clustering [1]-[3]. The method selection depends on the way the data are handled. Time-series clustering is Temporal-Proximity-Based Clustering if it works directly on raw data; Representation-Based Clustering if it works indirectly with the features extracted from the raw data, and Model-Based if it works with a model built from raw data, see [6]. The tested approach falls under the representation-based type and is expanded on by adding NLP elements in the feature extraction process.

II. UFO Data Ingestion and Analysis

We restricted our dataset to CY 2014-2015 and the US. Since the data are in html format, we imported it into Python Pandas using a web scraper. Some of these reports are spread across the United States, and hence, time-zone normalization is needed before any analysis could be performed. In order to have a better understanding of the UFO reports, we added the following external-source features: dates of astronomical events in CY 2014-2015, dates of national holidays in CY 2014-2015, US state population and share of active military population per year per each state. The resulting dataset contained 12,172 records. UFO sightings are displayed in a heat map (Fig. 1). The heat map revealed higher report density on the West Coast (California) and along the East Coast. Based on the report frequency distribution, the following states represent the top five in terms of reported UFOs: California, Florida, Washington, New York and Pennsylvania.

Harish Krishnamurthy (Data Scientist) is with the Zentek Masters LLC, Atlanta, India (phone: +1(617) 942-1729, e-mail: harish.kashyap@zentek.io).
Anna Lafontant (Data Analyst) is with the JEN Associates Inc, Cambridge, MA USA (phone: 867-321-3441, e-mail: annacogito@gmail.com).
Ren Yi (PhD Student) is with the Brown University, RI USA. (phone: +1 (631)561-7786, e-mail: renyi.ny2009@gmail.com).

A correlation matrix showed the expected logical correlations between the following features 'Week' and 'Quarter', 'Longitude' and 'Region West', 'Latitude' and 'Region South'. July and Friday are the month and day of the week with the highest number of reports, respectively, as shown in Fig. 3. The top five days with the highest number of reports are 2015-11-07 with 275 reports, 2014-07-04 with 203 reports, 2015-07-04 with 158 reports, 2015-07-05 with 91 reports and 2014-01-01 with 81 reports. Notably, two of the above dates are July 4th, one is New Years Day, and 2015-11-07 is the date of a confirmed navy missile test. We then tested the following hypothesis: Are there more reports on national holidays or days with astronomical events? We plotted the report frequency by day and colored the day of an astronomical event red and day of a national holiday green (see Fig. 3). The mean number of reports on the dates of a national holiday is 33 and the mean number of reports on the dates of an astronomical event is 16, which is same as the mean number of reports on the dates with no such events happening.

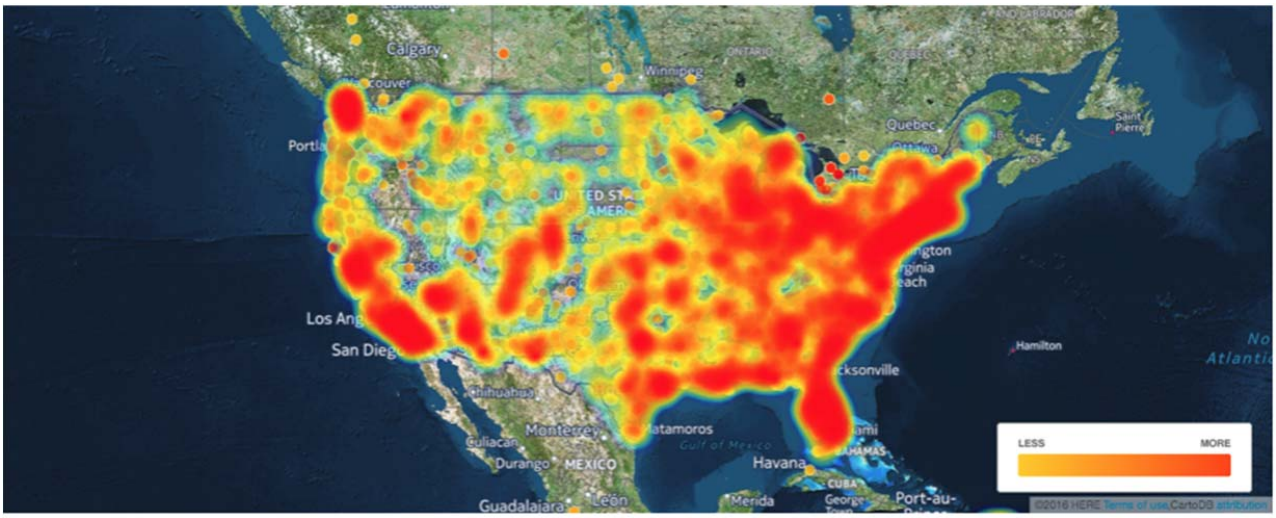

Fig. 1 NUFORC UFO reports in CY 2014-2015

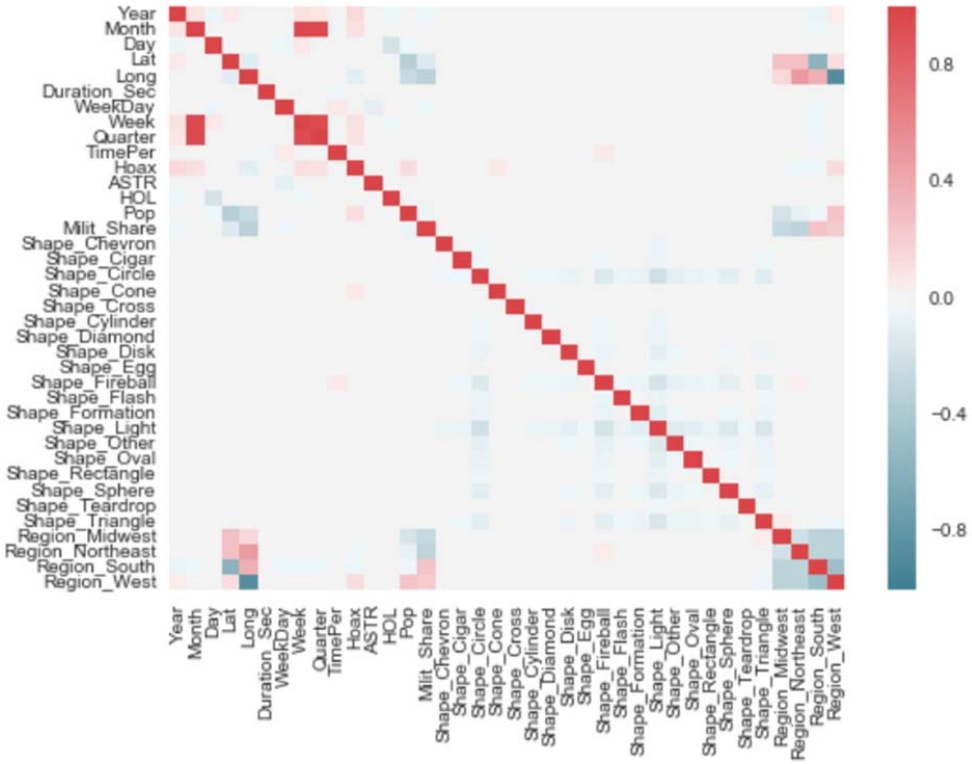

Fig. 2 Correlation matrix

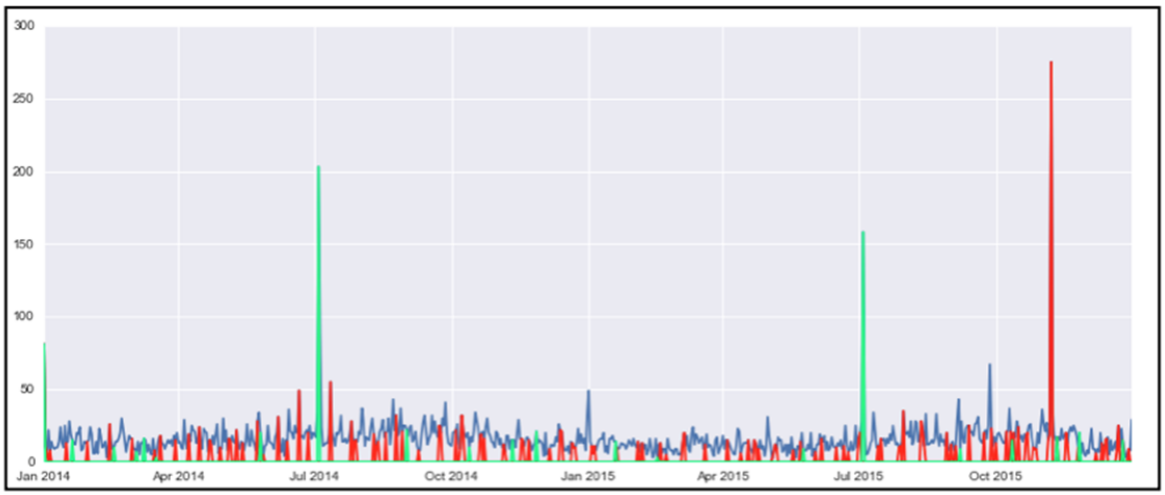

Fig. 3 UFO Report Frequency by Date CY 2014-2015: National Holidays vs. Astrological Events

Fig. 4 NUFORC Reports Heat Map with US Airports

III. Feature Extraction from Descriptions

NUFORC reports include a summary column, which describes the sighting. The descriptions are open-format and vary greatly in the level of detail. The average word count per entry is 14, with the maximum word count of 33. In the current section, we utilized the summary column for unsupervised semantic modeling in order to get some insight into what makes people think when they observe a UFO. We used Gensim - an open-source vector space modeling and topic modeling toolkit [3].

We used Latent Dirichlet Allocation model that represents documents as groupings of topic distributions and allows identifying top keywords with a given probability for each topic. The goal was to identify the most common topic among the NUFORC reports, where keywords would serve as pointers to the UFO perception question.

We apply this methodology in the following four steps:

1. Tokenization and Stemming/Lemmatization: We tokenized the words from each summary record by extracting them and removing morphological affixes from words, leaving only the word stem. We extract the common base form from stemming and lemmatization [4].
2. Vectorizing: We translated the set of words post lemmatization into a matrix, which is a sparse representation of the counts of words.
3. LDA model: When applying the LDA model, we specify the number of topics to be one, since our goal is to identify the top, most common topic. The number of key words was set to be five. Given that the mean count of total words (pre stemming and lemmatization) in the Summary column is 14, the first five are used as the keywords.

By computation, the five key words in topic 1 are 0.095 ∗ light + 0.047 ∗ sky + 0.030 ∗ bright + 0.030 ∗ object + 0.028 ∗ red.

The float component represents the probability of occurrence for a given word within a given topic. As part of our exploratory analysis, we tested if there are any correlations between four UFO reports locations and airport locations, since topic 1 is close to the description of night flights. Fig. 4 shows the heat map of the number of reports in the US, where the blue dots represent the location of all the airports. The map generally confirms the expectation that areas with a higher number of airports tend to report higher number of UFO sightings; however, the airport mapping also mirrors the overall population density and is not conclusive enough without further testing.

IV. Hoax Prediction

NUFORC reports contain those that are labeled as hoaxes by the internal team. We created a new binary column called Hoax based on the summary column and then tested if we can predict false reports based on other available features. The first model we used was the logistic regression model. The model did not perform well despite feature optimization efforts but served as an initial baseline estimate. See below for the ROC curve of the final model (0.499). Logistic regression was not the best model due to the non-linearity of the data.

The next tested model was random forest classifier. For parameter selection, we used cross-validation to determine the best values of n estimators, max depth and min samples leaf. Having started with all available features in the dataset, we used backward-elimination for determining the best features. The optimized model showed mean cross-validated ROC AUC score of 0.761 and mean accuracy score of 0.942. The following four features were the most important in terms of hoax prediction: Week, Region West, Time Period, and Duration. Despite high performance score of the model, the method has some limitations. In particular, the obviously false reports are deleted from the database (as opposed to being flagged as 'hoax'). One can make an argument that it allows the model to predict hoax reports that are not truly hoaxes but rather obvious cases of a different event being mistaken for a UFO (e.g., astronomical events, fireworks, planes, etc.).

V. Time Series Analysis

In the current section, we tested if the previous patterns in the UFO reports allow to predict the report count in the future using the following time-series methods: ARIMA model and Poisson regression. First, we considered the seasonal patterns in the reports. An ARIMA model predicted the number of reports. The procedures are the following: Split the original data into the train set and the test set, search for the best time window p and moving average q by assuming the differencing to be zero. After grid search for the best p and q of the smallest mean squared-error, we have p=3 and q=0. However, the performance of the model was poor. The predicted values were closer to the average count. As shown in Fig. 6, the blue plot represents the number of reports and the red plot is the prediction based on the ARIMA model.

In the next step, we used a Poisson regression to predict the number of reports in terms of time, see Fig. 7. The predictors for the Poisson regression are dates and dummy variables for national holidays and astronomical events. The Poisson regression performed slightly better, but still could not fit the time-series dataset. Hence, this shows the amount of randomness present in the data.

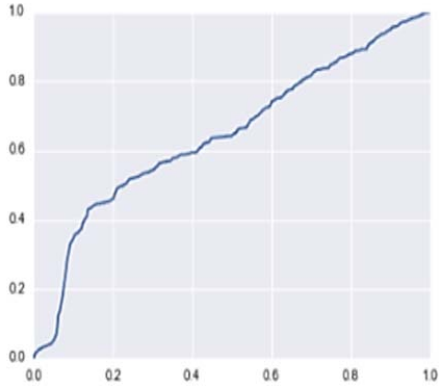

Fig. 5 Logistic Regression ROC curve

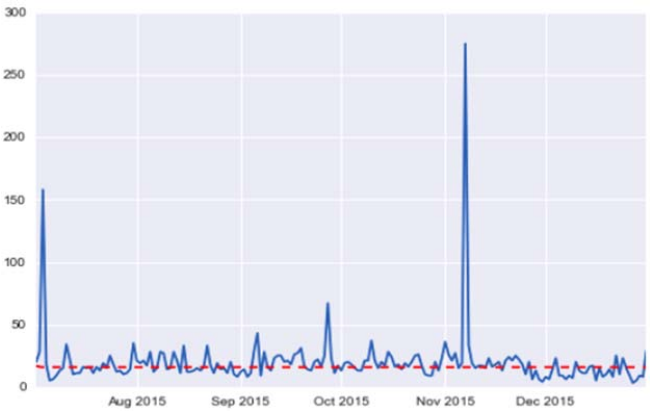

Fig. 6 ARIMA Model Prediction

VI. Feature Vectors for Clustering

It is important to extract features and form feature vectors for unsupervised clustering. This is to gain insight into what types of events might exist based on all the data we have. Here we assumed that descriptions containing several of the same words are characteristic of the type of event. For analysis, we looked at those days, where number of events peaked. One of the interesting days is the report on 11-07-2015. To form feature vectors and perform clustering, the top key-words are discovered by setting the number of topics to be three and for each topic, we look for five key words. To determine the number of topics parameters, the variable is varied between 1 and a large number such as 5, and the density distribution between clusters are examined to arrive a number as 3. We then gathered the unique set of keywords by aggregating all the topic keywords in no particular order. These were "bright",

"sky", "blue", "white", "green", "cloud", "bright", and "saw". The feature vector is then initialized with the key-words as unit vectors, with a 1 indicating a presence and a 0 indicating the absence of the former. Therefore, if the summary of the report contains the nth key word in the set K, then the nth entry of the vector is 1; otherwise the nth entry is 0.

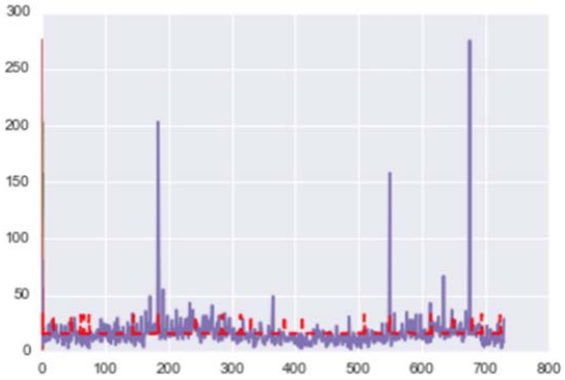

Fig. 7 Poisson Regression Model Prediction

Since the spatial data is also known, we capture that with the latitude and longitude information and append it to the feature vector. For unsupervised clustering, k-means was used for the vectors, where each vector, f(v) is represented as:.
f(v) = [0,1,0,0,0,0,0,1,lat,long]

VII. Unsupervised Clustering

For identifying the clusters amongst the dataset, we used k-means to generate cluster realizations. K-means suffers from the fact that it needs a number of clusters to be specified prior to running the algorithm. Hence, this necessitates searching the best value for the latter. There are many past methods available to determine the best cluster value. Since, we have a 2-D data set consisting of latitude and longitude that is appended with other features from the textual data, we could visualize various latitude and longitudinal results post clustering. This was especially helpful to determine the cluster density distribution across various clusters for each different realization. Since the data has time information, we consider an entire 24 hour window around the peaks prior to formation of feature vectors and clustering. The data needed processing such as conversion to a common time zone and parsing relevant information. This 24-hour window around the peak was gathered along with the latitude, longitude and description information to form features. Later, we searched for the optimum k to determine the cluster that contains relevant information about the event.

VIII. Results

Data on Nov. 7th 2015: Figs. 8-10 shows various cluster realizations for test events at their peak density. November 7th, 2015 had the highest number of reports in the dataset. A dense cluster was found centered in Southern California after performing clustering by varying the number of clusters parameters. Upon literature review and Google searches, it was discovered that there was a Navy missile test event that occurred from a submarine off the coast of southern California on that date. See Fig. 8 below for examples of cluster realizations at different stages.

Data on July 4th 2014: The data on this date is also one of the peaks in terms of density. However, due to the likely correlation with the fireworks the data points are spread across the country. The date had 49 qualifying data points, i.e., data points with latitude, longitude, date/time, and the summary present.

Counter intuitively, the data points in the upper left hand corner did not group into a standalone cluster. Based on the results, we can conclude that the lower part of the right side cluster is likely to be an event. Given that the identified closer points are located in the Texas area, it is possible that the event is related to NASA activity.

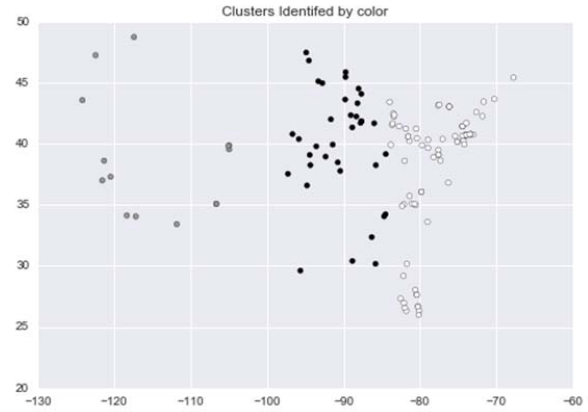

Fig. 8 Cluster Realization using three clusters

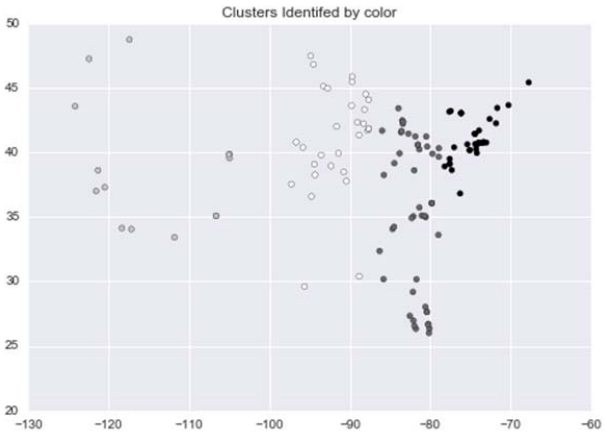

Fig. 9 Clustering Realization using four clusters

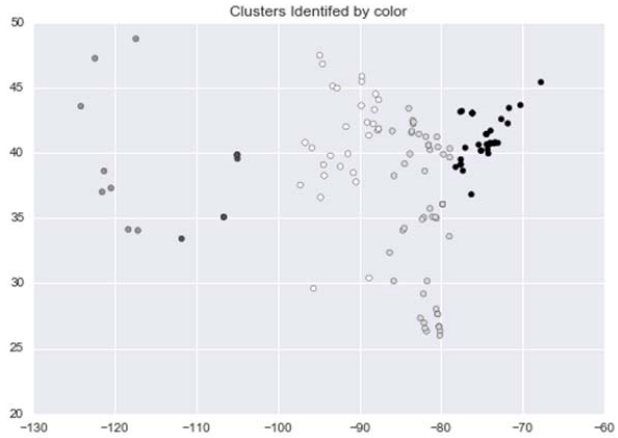

Fig. 10 Clustering Realization using five clusters

IX. Conclusion

The analysis revealed general trends in UFO reporting. In particular, the overall correlation between the report frequency, population density and the increased frequency in the summer months and weekends were high. The most predictive features in terms of hoax prediction were Week, Region West, Time Period, and Duration. We have been able to demonstrate spatial events in 2-D using latitude, longitude, and processed summary records with the k-means model. Future work will be focused on the method refinement and expanding the dataset for additional testing. The other areas to investigate are to use the time information and perform clustering to identify the relevant events.

Acknowledgment

This project started off as a course project for General Assembly and was continued after the course leading to interesting results. Hence, we thank the General Assembly for providing the opportunity.

References

[1] B. J. Goode , J. M. Reyes, D. R. Pardo-Yepez, G. L. Canale, R. M. Tong, D. Mares, M. Roan, N. Ramakrishnan. Time-Series Analysis of Blog and Metaphor Dynamics for Event Detection. Advances in Cross-Cultural Decision Making Volume 480 of the series Advances in Intelligent Systems and Computing pp 17-27.

[2] M. Moshtaghi, C. Leckie, and J. C. Bezdek. Online Clustering of Multivariate Time- series. Proceedings of the 2016 SIAM International Conference on Data Mining. 2016, 360-368.

[3] K. P. D. Artificial intelligence methods for theory representation and hypothesis formation. Computer Applications in the biosciences. 1991 Jul; 7(3): 301-8.

[4] N. Marin, D. Sanchez. On generating linguistic descriptions of time series. Fuzzy Sets and Systems Volume 285, 15 February 2016, pp 6-30.

[5] C. Rutkowski, G. Dittman. The Canadian UFO report: the best cases revealed.

[6] S. Rani, G. Sikka. Recent techniques of clustering of time series data: a survey. International Journal of Computer Applications (0975-8887) Vol. 52-No. 15, August 2012.