Distributing 5G Cell Sites Optimally using Spectral Clustering and K-Means Clustering

[1]Ajay Raj Nelapudi, [2]S.A. Bhavani, [3]Basa Ashok Kumar, [4] Vuke Sreeram Sagar, [5]Suryanarayana Durga Sai Krishna
[1]Student, [2]Assistant Professor, [3,4,5]Student
[1]Department of Computer Science & Engineering,
[1]Anil Neerukonda Institute Of Technology And Sciences, Visakhapatnam, India

Abstract : In order to support 5G, new cell sites have to be installed to handle millimetre waves. These cell sites need to be installed large in number compared to existing macro towers and should be backhauled properly. Distributing these cell sites across a given area or city can be quite difficult if done using manual methods. The approximate location of the users can be derived from their address. Using this geo-spatial data and clustering algorithms, we divide the users' locations into segments for installing base stations. Each of these clusters will then be grouped into further clusters to suggest appropriate locations and optimal number of cell sites required to cover all the users.

IndexTerms **- 5G network, Cell Site, Backhaul, Geo-location, Spectral Clustering, K-Means Clustering.**

1.INTRODUCTION

The introduction of 4G, mobile internet speeds have gone up by a large factor. Users started to transfer their videos or other documents over the internet instead of using flash drives because 4G has made the transfer faster and reliable. Many new users have joined the 4G network to avail these benefits while mobile network operators have competed to offer plans at lesser prices. This has led to overcrowding of the allocated 4G spectrum. Promising to overcome this problem and an increase in speed, 5G was developed and deployed to test in cities across USA. 5G is designed to work with millimetre waves. These waves can neither penetrate objects nor travel far. To accommodate 5G, a vast number of small cell sites have to be installed in a region where one or two 4G macro towers used to radiate waves. Fortunately, these cell sites are small, easy to install or maintain and cost much less than the previous generation towers. However, due to the large number of cell sites required, we need to distribute them efficiently to strike a balance between the number of cell sites required and area covered. In addition to this, we need to suggest appropriate locations for the base stations which is where the cell sites are backhauled.

2.RELATED WORK

In 2014, Ms. Prukalpa Sankar and her team at Social Corps used geolocations and clustering techniques to help the government establish 10,000 LPG centres all across India. Inspired by her act, we set ourselves to see how can we apply such data mining techniques to build a better mobile network.

Since 5G and its associated problems have appeared very recently, there isn't much research done directly addressing the 5G radiation range problems or towers distribution. However, we will be able to use the foundations laid by other researchers for planning an effective network. Jocelyn Edinio Zacko Gbadoubissa,

Ado Adamou Abba Ari and Abdelhak Mourad Gueroui have proposed an approach to manage cell sites using kmeans clustering algorithm. Their approach proposed an initialization scheme that would reduce the sum of squared errors, SSE while providing a faster convergence. The SSE is an important factor not only in calculating the error but also in determining the number of clusters in a given dataset. We can calculate the mean squared error, MSE from SSE. The MSE will be useful in our case since we can have a fixed value for maximum permissible MSE. Also, because of the initialization scheme, we not only achieve local optima but also globally. This means that the location of each data point is at the least possible distance to its corresponding centroid while the regions covered by each centroid are in a state of least overlapping. These properties enhance the efficiency of our model.

All these cell sites should be backhauled to base stations. These base stations not only connect the cell sites to the nation's internet backbone but also helps when a user transfers from one cell site to another. Ideally, base stations should allow expansion of cell sites with the growth in users. This means that the number of base stations required cannot be calculated by the number of suggested cell sites or their locations. Shuai Yuan, PangNing Tan, Kendra Spence Cheruvelil, Sarah M. Collins and Patricia A. Soranno have proposed a constrained spectral clustering algorithm for regionalization. Constrained spectral clustering uses domain knowledge while clustering to find appropriate clusters. However, the balance between spatial contiguity and landscape homogeneity should be maintained. This is done by using a spatially constrained kernel matrix. Then a truncated exponential kernel is computed from it, which is then binarized to be represented as an adjacency matrix. Using this, we will be able to identify different regions in the geo-spatial data of the users' locations. Each region can have a base station that the cell sites can connect as backhaul. This way there is room for expansion for future purposes.

While spectral clustering helps in identifying regions, it does not tell us the optimal number of clusters in a given dataset. This means the no of clusters should be fixed before applying the clustering techniques into which the entire dataset is classified as. However we are required provide an optimal number of regions to decide the base stations. Lihi Zelnik-Manor and Pietro Perona proposed a self tuning spectral clustering. An affinity matrix is first computed and then normalized. Their Eigen values and Eigen vectors are calculated. Then the maximum gap is identified which corresponds to the number of clusters by Eigen gap heuristic. They have found that the clusters are now optimal considering that they used a local scale which is estimated from the data set itself.

3. METHODOLOGY

The first phase involves dataset preparation. In our case, we considered the area of Visakhapatnam City. The format of each data point is a latitude and longitude put in a CSV file. In real time, say a mobile network operator wants to use this while planning to launch 5G, he needs the users' locations as well. The problem is that the users do not provide their co-ordinates while filling up the form for applying for a connection. However, they fill in the address string. Using google maps geocoding API, we can convert each of the users' address string to its corresponding geo-coordinates. This data can be fed as input to our model to obtain the desired result. In our case, we used our knowledge of the city to locate and manually plotted and saved them to a hotspots.csv. Then we used this file and scripting to randomly generate more points around each hostpot and saved it as dataset.csv. The former file hotspots.csv helped us produce different datasets during evaluation phase.

Our choice of language is python because of its vast libraries that support numerical and scientific computations in addition to data mining algorithms. Using these libraries we built the affinity matrix upon the gathered dataset. Upon this matrix we calculate the graph laplacian and thereby the Eigen values and vectors. Once computed, the maximum gap between the Eigen values is identified and considered as a good number of clusters. Using this value, we apply spectral clustering to identify the regions. The result is shown in figure 1. Each region will have a base station that the cell sites can backhaul to. Spectral clustering does not have centroids. Though we could calculate the centroid explicitly, it won't be an appropriate place to install a base station since this clustering technique can identify different shapes and the centroid may not always be amongst the dataset. For instance, we might have identified a region with a human settlement around an elliptic water body. The centroid, if calculated, would be in the water. This is not an ideal case. However, the centroid of a cluster is useful in further calculations.

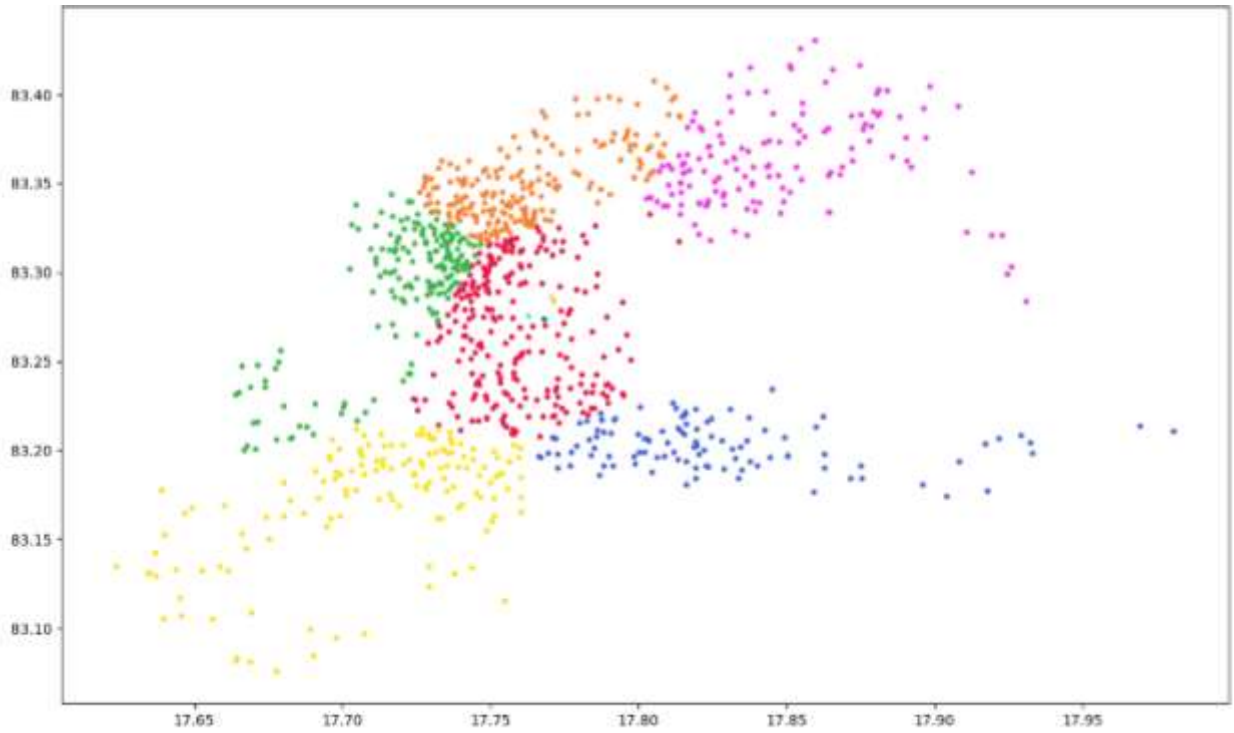

Figure 1: Result after Applying Spectral Clustering

For each of the region identified above, we need to distribute cell sites. For this purpose, we use kmeans clustering. The centroids thus obtained could be the locations of cell sites. However, we also need to propose the number of clusters for optimal division of the region amongst the cell sites. The MSE determines the mean distance from each user to its corresponding cell site. In our case, the MSE should not exceed 0.0064 which is the square of 0.08, where the latter value is the difference between any two geo-coordinates which are 800 metres apart. This is because the average range of a millimetre wave is about 1000 metres or 1 kilometre. However, fixing this parameter to 1 kilometre would disturb the global optima, which means the users will find a drop in signal when they move from cell site to cell site. We iterate like the elbow method until the MSE is less than 0.0064 to decide the number of optimal clusters and their centroids. Thus, the obtained centroids will be locations of cell sites. This can be seen in figure 2.

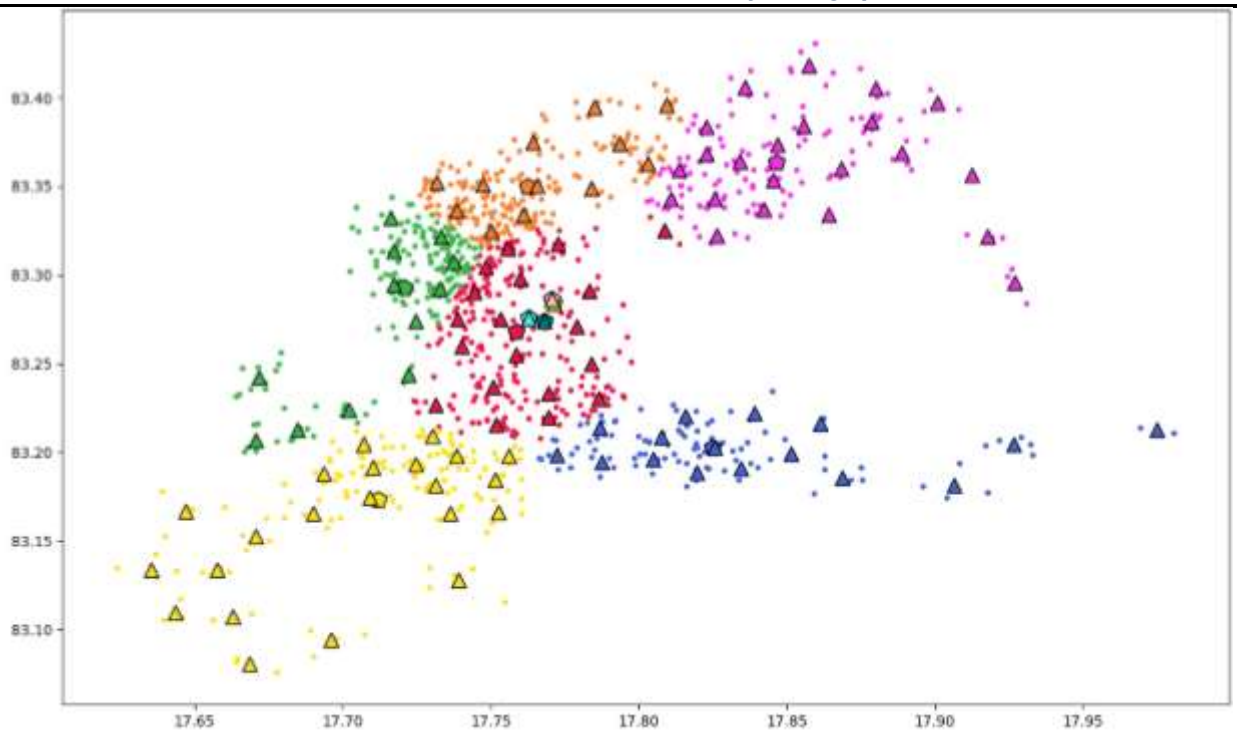

Figure 2: Distribution after Applying K-Means Clustering

However, given the shape and orientation of the geo-spatial dataset of the users' locations, it might be possible for our model to generate clusters with one base station and one cell site. This is not an ideal case and will decrease the efficiency of distribution with increase in cost of installation. This phenomenon can be seen in figure 1 where 4 new clusters appeared within the first cluster. This will lead of suggesting a single base station and a single cluster as seen in figure 2. To help reduce this problem, we require certain adjustments. Any region with less than 5 cell sites is clubbed with a region closest to it. Here the region centroid is useful. Comparing the current centroid to other region centroids help in determining the region closest to it.

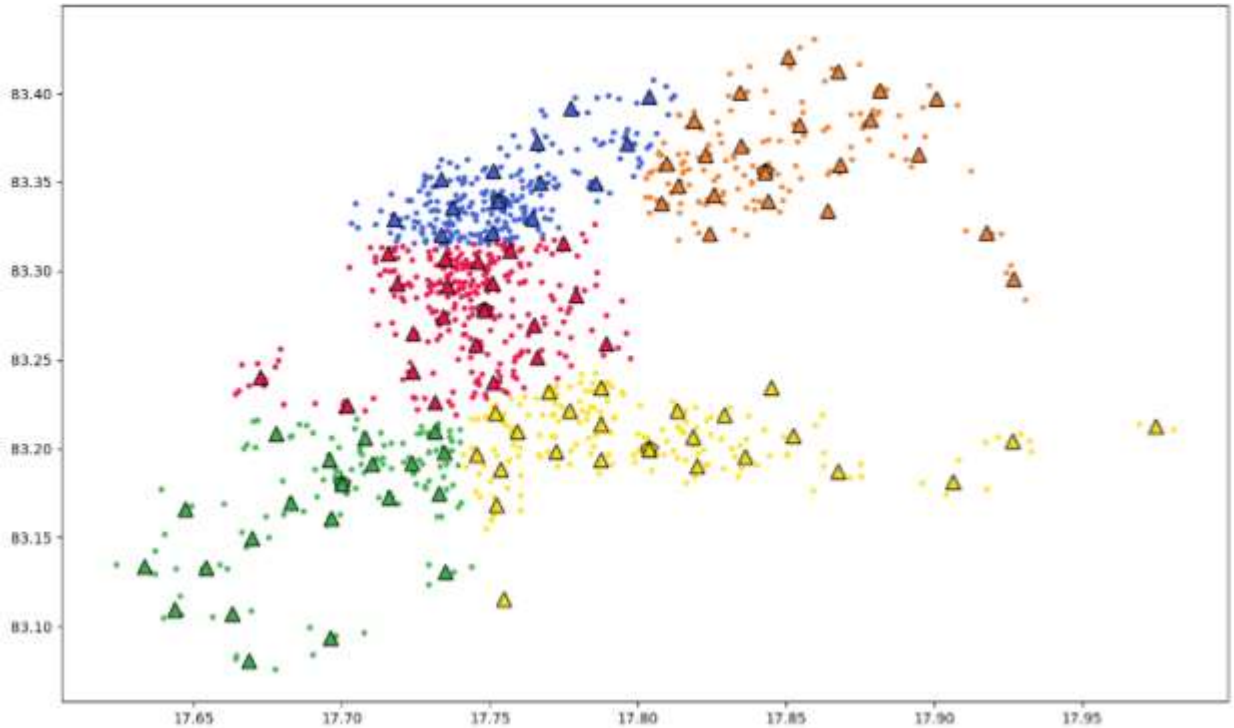

Figure 3: Distribution after Adjustment Phase

When clubbing the regions, any cell sites found within the range of 300 meters from each other are removed to reduce overlapping area of cell site radiations. This method ensures that both the local and global optima are preserved. The cell site closest to the region centroid is considered as the base station. This is because a cell site is always amongst the data points and so will be the base station. A proof of this adjustment can be seen in figure 3.

```
{
        str([base_station_lat, base_station_lon]):
                {
                        'base_station': [lat, lon],
                        'cell_sites': [[lat1, lon1], [lat2, lon2], ..., [latN, longN]],
                        'users': [[lat1, lon1], [lat2, lon2], ..., [latN, longN]]
                }
}
```

Figure 4: Tower Distribution Data Structure

Our distribution requires an efficient data structure to store and the results. This structure should be serializable and deserializable because the results of our software might be taken as input for another tool. Hence, we decided on the data structure shown in figure 4. Since each region will have only one base station, we decided to stringify its coordinates and use it as the key. Its value is another dictionary consisting of base station, cell sites and its corresponding users in Numpy array format. This structure is serialized to a JSON format and saved in a file with ".json" extension.

1. RESULTS AND ACCURACY

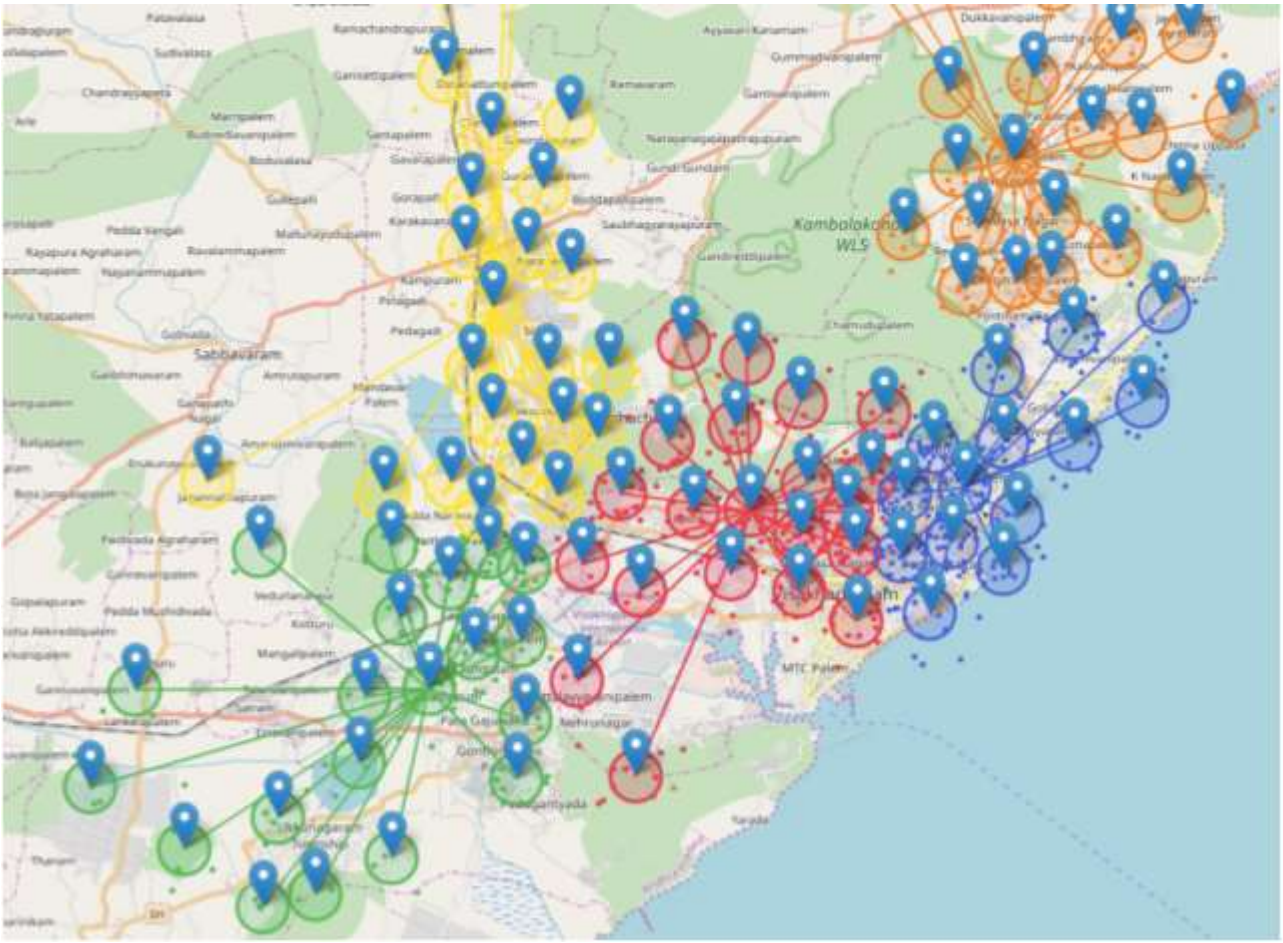

Figure 5: Distribution Visualized using Folium

Using folium, we display the results of our distribution. For effective visualization, we represent each region in a distinct colour. The cell sites are represented as blue markers while the base stations are shown as black tinted circles with its corresponding coloured border. Lines are drawn from each base station to all the cell sites in its region, in its relevant colour. This navigable map is generated in HTML format and is opened using the default web browser to view the result. Since, the true intent of our visualization method cannot be viewed on a paper, we hosted our sample output at https://towersdistributor.000webhostapp.com/. However, a screenshot of the result is shown in figure 5. The overlapping of cell site zones shown in Figure 2 is diminish viewed at a proper zoom in the browser.

Since we know the range of millimetre waves and we have the locations of cell sites and the users, we evaluated our result by calculating the ratio of users who fall within the range of at least one cell site to the total number of users. As mentioned earlier, we use the hotspots.csv for evaluation. We generated different datasets atop the hostspots file and applied our model. An unrealistically sparse dataset when fed to our model provided an accuracy of 0.7635 which is excellent considering that the users are spread across a large area and are placed far from each other. However running over model over a sample of 20 different datasets with a realistic density of the city gave us an average accuracy of 0.9310. It is to be understood that we can tweak our model to provide a higher accuracy than the current result but the result would be unrealistic. Meaning, that operators will consider their return on investment ratio and discard certain cell sites. Removing only few cell sites and not adjusting the others will disturb the global optima which in turn will reduce the accuracy by a significant factor.

FUTURE SCOPE AND CONCLUSION

Our method distributes cell sites optimally all across the given users' locations. We also understand that each cell site is specific to a set of users and we might need to add more cell sites. Keeping this in mind, we distributed the base stations to allow expansion of the network as well. This means the distribution is not only optimal or reliable but is also future proof. Furthermore, we can make changes in the model to allow much appropriate distribution by considering certain features of the land such as elevation. This modification will allow the model to give more accurate results in hilly regions. It is quite possible that in the future users might have the option to allow the mobile mobile network operators to host 5G cell sites atop their properties just like how 4G towers are installed. This means that our model could be tweaked with a few fixed centroids and the others should be clustered. Though our model distributes the cell sites with a high accuracy, the mobile mobile network operators may not install all these towers based on the budget and return on investment. Hence, building a business model atop this and showing insights of profits could help make better decisions. While these are the wide variety of features that can be incorporated into this model to make it more customizable, we believe our approach to distribute 5G cell sites optimally will lay the foundations to future work. Also, by providing this solution prior to a full 5G network deployment, mobile mobile network operators can use this model to plan and build their 5G network accordingly.

REFERENCES

[1] Shuai Yuan, Pang-Ning Tan, Kendra Spence Cheruvelil, Sarah M. Collins and Patricia A. Soranno(2015) "Constrained spectral clustering for regionalization: Exploring the trade-off between spatial contiguity and landscape homogeneity" IEEE International Conference on Data Science and Advanced Analytics (DSAA).

[2] Jocelyn Edinio Zacko Gbadoubissa, Ado Adamou Abba Ari and Abdelhak Mourad Gueroui(2019) "Efficient kmeans based clustering scheme for mobile networks cell sites management" Journal of King Saud University.

[3] Lihi Zelnik-Manor and Pietro Perona(2004) "Self Tuning Spectral Clustering" Advances in Neural Information Processing Systems (NIPS).

[4] Francky Fouedjio "A spectral clustering approach for multivariate geostatistical data" International Journal of Data Science and Analytics 4:301–312 DOI 10.1007/s41060-017-0069-7

[5] Lamiaa Fattouh Ibrahim and Manal El Harby(2012) "Enhancing Clustering Algorithm to Plan Efficient Mobile Network" International Journal of Computer Applications (0975 – 8887) Volume 59– No.18.

[6] Arvind Sharma and R.K. Gupta(2017) "Spatial Data Mining with the Application of Spectral Clustering: A Trend Detection Approach" International Journal of Computer Applications (0975 – 8887) Volume 173 – No.2.

[7] K.Kameshwaran and K.Malarvizhi(2014) "Survey on Clustering Techniques in Data Mining" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2272-2276. variables as common pervasive risk factors and the empirical content of the Arbitrage Pricing Theory. Journal of Empirical finance, 5(3): 221–240.