Defining Covid 19 containment zones using K-means dynamically

Dr. Satish Chinchorkar (chinchorkar@gmail.com)
ValueBoosters https://orcid.org/0000-0001-7138-6586

Method Article

Keywords: Covid-19, K-means, quarantine, hot-spot zones, contentment, clusters, Google Map

Posted Date: September 25th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-83392/v1

Abstract

In view of the rapid growth of Covid-19 pandemic, contagious nature of the disease and non-availability of effective vaccine; the only way available is to restrict the people's movements from mixing in a mob. However imposing total lockdown may not be the feasible solution because it is not only counter-productive but also causes the destructive impact on day-to-day working, economy and convenience of people. Moreover total lockdown is at the cost of public freedom may cause people agitation. Therefore determining the micro-level, manageable quarantine zones for affected Corona positive patients and further focus to only on the identified zones can be the resolution. For this purpose the scope of the containment zones must be determined with unbiased, precise and agile manner to enforce the controls on these zones to prevent the spread of this contagious disease. The updated and accurate information about such hot-spot zones can be useful for government to effectively implement the measures by concentrating the efforts on the zones and for other citizens to alert such hot-spot zones. However the task of identifying and circumventing the precise affected zones is not easy because of the constantly changing status of the patients. As soon as number of patients are getting recovered (the cycle time is around 14 days), these quarantine zones need to be revised and reconfigured accordingly, which is in addition to constantly accumulation of the data of new patients. The size and locations of such zones (affected by Corona positive patients) is dynamic in nature, therefore it becomes impossible to frequently reconfigure it manually. Implementing the models such as K-means from Data Science is proposed to help the situation because the zones determined by Data Science models are reliable (fact-based and latest), economic (not much additional infrastructure required), easy to understand (clusters are well defined and visible), flexible (can be parameterized / configured), and unbiased (because there is no preconception while defining zones/ clusters).

Introduction

"COVID-19 is a Data Science issue" (Callaghan, 2020) the comprehensive article gives various ideas and inspiration to think about the data and how it can be effectively used in current pandemic situation.

Quarantine is nothing but the separation and restriction of movement or activities of persons who are not ill but who are believed to have been exposed to infection, for the purpose of preventing transmission of diseases. Persons are usually quarantined in their homes, but they may also be quarantined in community-based facilities. Considering the increasing volume of number of patients and limited community-based facilities most of the people are being asked to quarantine in their homes.

The Cluster Containment Strategy would be to control the disease within a defined geographic area by early detection of cases, breaking the chain of transmission and thus preventing its spread to new areas. This would include geographic quarantine, social distancing measures, enhanced active surveillance, testing all suspected cases, isolation of cases, quarantine of contacts and risk communication to create awareness among public on preventive public health measures. Many sensitive factors are associated

while defining the containment zones. The demarcation should be unbiased and latest because many funds and government aids as well as restrictions are directly related to definition of such zones.

Manually defining these constantly changing clusters in size and location may not be feasible. The techniques from data science specifically K-means can help in the situation by defining the clusters as containments.

Significance

Considering the rapid rate of outburst of COVID-19 and the fact there is no proven vaccine available yet, only prevention can be possible. Though bitter but prohibiting the people mixing in the mass community is the only control as the nature of this disease is contagious in nature. However declaring state-level or district-level lock-down is not advisable. Total lock-down (without containment zones) causes not only inconvenience to the citizens but it result in tremendous setback to running economy. Hence strategy of defining micro-level contentment zones and control or manage them effectively is recommended. This shall help to have balance between lock-down and keep the industries / business running.

However the information of patients (specifically about location) is mutable, voluminous and constantly changing hence application of Data Science Models are recommended. Moreover the visibility and reliability of solutions achieves the objective.

Objective And Scope

In traditional way planning, executing, maintaining this information and generation the useful output is just not feasible. Deploying the advance technology and techniques like Data Science becomes vital.

The objective of this paper is to bring the objectivity and accuracy in creating the COVID-19 containment zones dynamically and consistently using K-means technique of Data Science. This is needed to get continuous updating on accurate and latest micro-level demarcations information of contentment zones to plan the unbiased strategies for separating contacts of COVID-19 patients from community.

The scope of this paper is restricted to provide the concept to apply the K-means technique from Data Science on the collected patient's geographical data like locations to define and visually plot the contentment zones (clusters) on the actual map.

Literature Review

4.1 COVID-19 and Containment

According to (Wollersheim, 2020) during the COVID-19 crisis the field of Data Science is in center. Most of the community is interested, watching and looking forward the statistical analysis and epidemiology graphs and sharing the same in social media on a large scale. The expectation from Data Science is very

high. Data Science is emerging field consist of number of applicable and useful tools, techniques and functions, using which taking the fact-based decisions and planning can be possible, which is very essential in current situation.

The cluster containment strategy for Zika virus outbreak (Singh et al., 2019) was found effective in Rajasthan, India. Singh et al (2019) in their paper explained that how surveillance strategies used to control the disease from spreading beyond containment zones of 3 km radius. The article gives emphasis on creating to containments to prevent the outburst of disease, however it does not explain about how to make these zones quickly and accurately. In their paper (Maier & Brockmann, 2020) explained about the effective containment to control specifically COVID-19 cases in China. The model which they explained in their paper captures both quarantine of symptomatic infected individuals and other population isolation practices. The focus of the research is on contagion process and general effects as well as significance of the containment. Their research work implies and supports the need to define the containment zones accurately.

As stated in old article of Teena (2020), the Government of India had given a broad guidelines to classify the containment zones in three types as Green-Zone (if there are no confirmed cases or no report of cases since last 21 days), Orange-Zone (where zonal retractions can be relaxed based on situation) and Red-Zone (containment zone where strict lock-down can be imposed). As per Teena (2020), Government asked district administration to demarcate the containment areas with red and orange zones around connection with the Coronavirus outbreak boundary of containment zones as colony, mohalla, ward and police station area etc. Which support the need to micro-level defining and updating the containment zones.

4.2 Application of Data Science

A comprehensive review of application of Data Science to combat COVID-10 (Latif et al., 2020) explains how in general the Data Science is going to play central role to control the pandemic. This paper broadly covers the features of Data Science in tracking the spread and mitigation strategies that includes identification and resolving the gaps in surveys and updating the records. The paper addresses use of Artificial Intelligence (AI) for risk assessment and prioritization, use of mobile applications (such as Babylon) for screening and diagnosis. Simulation techniques used to reduce the impact of pandemic. Application of Data Science used in contact tracking, logical planning and social / economic intervention. Automated patient care, supporting vaccine discovery is supported with use of techniques such as Auto Regressive Integrated Moving Average (ARIMA), Long Short Term Memory (LSTM) and Latent Dirichlt Allocation (LDA). The paper equates the correctness of results v/s urgency. Many systems proposed in the paper are not operational.

The application of K-means clustering, the unsupervised method of data analysis using constrained with background knowledge (Wagstaff et al., 2001) explained by imposing the GPS data to resolve the real-world road-lane finding problem. They used COP-KMEANS and COP-COBWEB methods. This article gives idea about assignments of instances of clusters.

The article (Weatherill & Burton, 2009) explains how the seismic source zones were created using K-means cluster analysis for the Aegean region. The paper describes the significance of applying K-means algorithm for hierarchical cluster analysis was found useful for partition regions based on observed seismicity to have consistent approach to source model development. Two techniques were adopted as 1. Point Source K-Means: used to partition a catalogue of earthquake hypocenters and 2. Novel Line Source Development: Line K-Means algorithm is used for partitioning a set of line segment using catalogue of known fault ruptures. The challenge of finding the appropriate number of clusters is resolved by 'cluster quality index' technique used to identify the optimum number of clusters.

Cluster Analysis classifies the data-points (locations) in to groups (such as Green-zone, Orange-zone and Red-zone) based on set of variables (similar to Factor Analysis) such as whether patient is positive or not. Three major steps are involved in conducting the Cluster Analysis a) Choose a similarity measures b) Choose a clustering procedure (K-means, in this case) and c) Decide on number of clusters.

The clustering procedure which is the core part of this paper, are broadly classified into hierarchical and non-hierarchical procedures. In Hierarchical clustering the cluster memberships of objects (data-points) are involved through a step-by-step hierarchy or treelike structure, this type of clustering is further classified in to Agglomerative and Divisive clustering. However for the purpose of making the containment zones Non-Hierarchical clustering is being considered which also known as K-means is clustering. There are three methods in this clustering a) sequential threshold b) parallel threshold and c) optimizing partitioning. The sequential method is recommended where the reassignment of cluster membership is not allowed. i. e. Once the object (location-data-point) is assigned to a cluster, it will not be considered for further analysis of reassigning it to another cluster at a larger stage. This procedure need to have pre-determination of number of clusters and will give arbitrary selection of cluster centers.

Proximity Matrix (Refer Annexure-C) is prepared which gives the squared Euclidian distances between each pair of the cases (containment zone location here).

Hypothesis

Applications of K-means technique of Data Science contribute towards defining, visualizing and maintaining the containment zones impacted by COVID-19.

Proposed Design And Methodology For Application Of The Data Science

6.1 Methodology

The paper is conceptual in nature and is based on the personal experience, interviews and study of past research conducted. A systematic approach is proposed to define the containment zones objectively (based on facts/data). This includes the data-collection and data-processing phases with suggested tools and techniques. As shown in Fig.1 the location data of patients is being collected using app such as

SAHYOG on mobile devices. The data collected in central data base shall be preceded further for creating the clusters (using K-means cluster analysis) and further imposed on the location maps.

Data Collection

The mobile application SAHYOG, as well as the web portal (https://indiamaps.gov.in/soiapp/) prepared & managed by the Survey of India (SoI), has been customized to collect COVID-19 specific geospatial datasets through community engagement to augment the response activities by Government of India to the pandemic. Information parameters required as per the Govt. of India strategy and containment plan for large outbreaks have been incorporated in the SAHYOG application.

Data Processing

As shown in Fig.2 Visualization of within-cluster quantity indices: (A) total within cluster sum of squares, (B) pooled within-cluster sum of square distances.

Application of K-means clustering for COVID-19 (Siddiqui et al., 2020) which is distance-based, fast-processing and has linear complexity O(n)[13]. Simple four steps are involved to determine the clusters as follows:

Step-1: Select the number of ‘K’ clusters to be identified, represent in the initial group of centroid. This selection can be based on convenience of local Government administration bodies. The number indicates the proposed number of containment zones that can be managed based on available resources and infrastructure.

Step-2: Measure the distance between that point to each group centroid and classify the point closest center to it accordingly.

Step-3: Recalculate the group centroid based upon classified points. The logic of K-means is applied here.

Step-4: Repeat the Step-2 and Step-3 until the centroid does not change.

Refer Annexure-A for the system flow-chart of proposed model.

6.2 Discussions

Application of appropriate software to determine the clusters and plot on the graph is recommended. The R function provides kmeans() gives the algorithm for determining the clusters after installing the stats package. The clusters can be further plot using ggplot() and ggmap() functions available in R.

K-means provides two major types of clustering soft-clustering and hard-clustering. Hard-clustering is proposed as it shows each data point / object belongs to cluster or not. Clusters are formed based on four kinds as Connectivity-based clustering, Centroid-based clustering, Distribution-based clustering (where Gaussian algorithm is used) and Density-based clustering. Density-based clustering is proposed

as while clustering higher density within data space is considered and other data points are ignored treating them as noise. DBSCAN and OPTICS can be used to define the cluster borders. Refer Annexure-B and Annexure-C for algorithm.

Findings And Recommendations

The open source tools like R found useful for exploring the appropriate and applicable functions in the data science.

Several partitions with different values of K (number of clusters / partitions) are recommended to review along with cluster quality index for optimum solution.

To achieve the lock-down as well as keeping the industries / business running together it is recommended that the Data Science techniques such as K-means can be adapted to micro-level demarcation of containment zones.

Conclusion

To achieve the golden balance between lock-down as well as keeping the industries / business running together it is recommended that the Data Science techniques such as K-means can be adapted to define the micro-level demarcation of containment zones and manage them effectively. The clusters formed based on COVID-19 patient's locational data using Data Science techniques (specifically K-means) will be agile, unbiased, accurate, visible, economic and easy to apply.

Further Challenges On Large Outbreak Using Proposed Cluster Containment

Though application of Data Science models are useful for dynamically demark accurate and latest cluster containment, to control the large outbreak further challenges are still there as follows

(i) Deciding the number and size of the cluster/s.

(ii) Effectiveness of geographic quarantine.

(iii) Other environmental factors especially temperature and humidity and centers should be considered.

(iv) Public health response in terms of active case finding, testing of large number of cases, immediate isolation of suspect and confirmed cases and quarantine of contacts.

(v) Geographical characteristics of the area (e.g. accessibility, natural boundaries).

(vi) Population density and their movement (including migrant population).

(vii) Ability to ensure basic infrastructure and essential services.

(viii) Correctness of patient data with locational accuracy with facility of central secured database.

References

Callaghan, S. (2020). COVID-19 Is a Data Science Issue. In *Patterns* (Vol. 1, Issue 2, p. 100022). https://doi.org/10.1016/j.patter.2020.100022

Latif, S., Usman, M., Iqbal, W., Qadir, J., Manzoor, S., Tyson, G., Castro, I., Razi, A., Kamel Boulos, M. N., Weller, A., & Crowcroft, J. (2020). *Leveraging Data Science To Combat COVID-19: A Comprehensive Review TESCON (Tools for Enforcement of Smart Contracts) View project CONTRIVE View project Leveraging Data Science To Combat COVID-19: A Comprehensive Review*. *April*. https://doi.org/10.13140/RG.2.2.12685.28644/4

Maier, B. F., & Brockmann, D. (2020). Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. *Science*, *368*(6492), 742–746. https://doi.org/10.1126/science.abb4557

Siddiqui, M. K., Morales-Menendez, R., Gupta, P. K., Iqbal, H. M. N., Hussain, F., Khatoon, K., & Ahmad, S. (2020). Correlation Between Temperature and COVID-19 (Suspected, Confirmed and Death) Cases based on Machine Learning Analysis. *Journal of Pure and Applied Microbiology*, *14*(suppl 1), 1017–1024. https://doi.org/10.22207/jpam.14.spl1.40

Singh, R., Gupta, V., Malhotra, B., Singh, S., Ravindran, P., Meena, D., Gupta, J., Mathur, V. K., Mathur, R. P., Singh, S., Sharma, P., Sharma, H., Bhandari, S., Gupta, N., Sapkal, G., Mourya, D. T., & Speer, M. D. (2019). Cluster containment strategy: Addressing Zika virus outbreak in Rajasthan, India. *BMJ Global Health*, *4*(5), 1–4. https://doi.org/10.1136/bmjgh-2018-001383

Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001). Constrained K-means Clustering with Background Knowledge. *International Conference on Machine Learning ICML*, *pages*, 577–584. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.4624&rep=rep1&type=pdf

Weatherill, G., & Burton, P. W. (2009). Delineation of shallow seismic source zones using K-means cluster analysis, with application to the Aegean region. *Geophysical Journal International*, *176*(2), 565–588. https://doi.org/10.1111/j.1365-246X.2008.03997.x

Wollersheim, B. C. (2020). Surprising Side Effect of COVID-19 : We are All Data Scientists Now. *Data Analytics & Insights, Arcadis*.

Figures

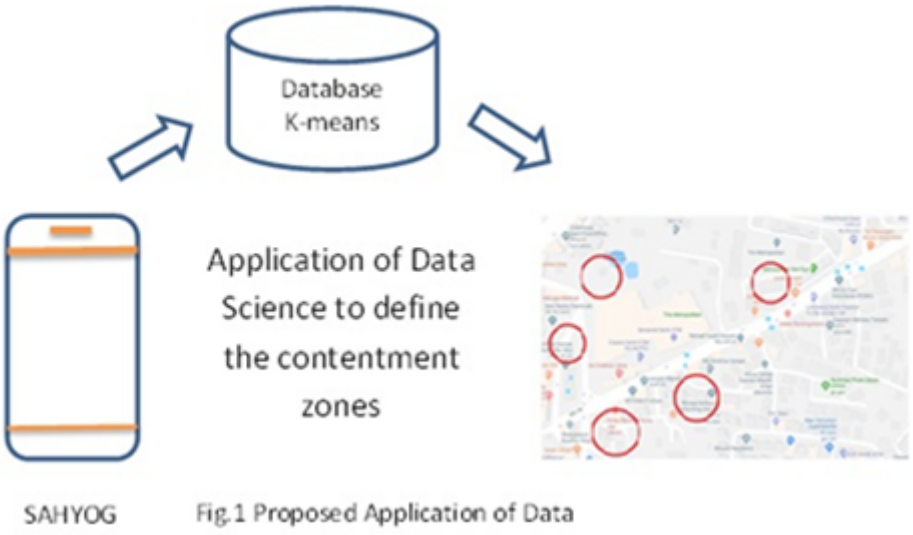

Figure 1

Proposed application of data

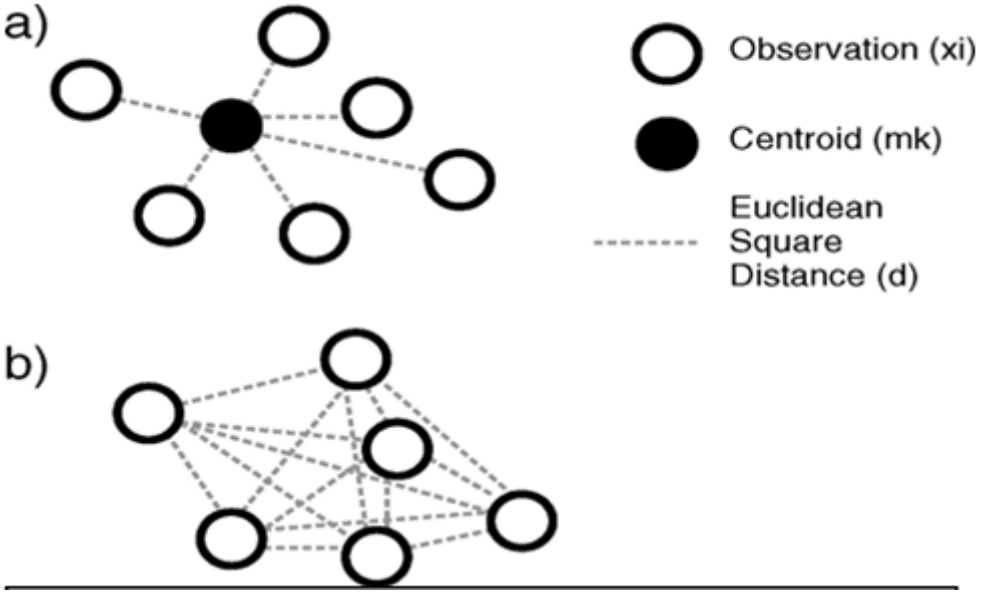

Figure 2

Visualization of within-cluster quantity indices: Ref. Volume 176, Issue 2, February 2009, Pages 565–588, https://doi.org/10.1111/j.1365-246X.2008.03997.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx