International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Analyzing the Epicenters of the Coronavirus in the U.S. with Modified K-Means Clustering Technique

Jonathan Tri Pham[1]
[1]Department of Natural Sciences, University of Houston - Clear Lake, Houston, TX 77058

Abstract— *With the Coronavirus pandemic posing a serious threat to human lives, the daily listing of confirmed cases and deaths is a source of information indicating the severity of the crisis. However, this daily listing is difficult to digest due to the enormous amount of numerical data involved, reducing the casual readers to a quick glance at the areas with the highest case counts with no further critical analysis. This paper presents the use of modified K-Means clustering technique to identify critical geographical areas surrounding the epicenters of the pandemic. Attention and effort can be directed to these hotspots in order for the inevitable effects of the pandemic to be alleviated through carefully planned governmental policies and the corresponding compliance of the citizens.*

Keywords— *Coronavirus pandemic, epicenters, modified K-means clustering, model based initialization, semblance method.*

I. Introduction

The Coronavirus (COVID-19) pandemic that is currently happening in 2020 is a global crisis affecting the lives of everybody around the world [1]. Within the United States, the occurrence has been reported for each of the 3143 counties and updated on the daily basis since January of 2020, and this data set is maintained in a database that is made openly available to interested parties [2]. However, the data set is large and contains too much data at the detailed numerical level that makes it tediously difficult to read and analyze [3]. The phenomenon of having too much information is often seen as an impedance to meaningful and timely analysis of the available data, defeating the purpose of gathering data and making it available [4].

The role of information technology, designed to address the common case of not having sufficient information, has always been to provide relevant information in a timely manner to support the decision process that results in some impacting manner [5]. However, with the advance of information technology, the problem of not having sufficient information has been solved in most common scenarios [6]. One unintended consequence of information technology is having too much information at hand that relevant and useful information becomes buried by itself to the point of being unobservable or unnoticeable by human beings [7]. In this extreme scenario, a large quantity amount of information needs to be analyzed and meta-information needs to be extracted so that key points can be identified for human consumption at first glance [8].

Data mining is a discipline that uses computers to automatically analyze large sets of data and extract information that human normally and perhaps routinely missed [9] due to the human nature of seeing things in a focused manner that is often abstractly described as tunnel vision [10]. The process of data mining consists of four sequential stages: segmentation, representation, compaction, and classification [11]. The first stage of segmentation that sifts through data to identify clusters of data with similar characteristics is sometimes (subjectively) considered the most important stage because it affects the relevance of the results of the other remaining stages that must depend on the results from the previous stages [12]. In this aspect, it is important to recognize various clustering techniques of data segmentation, and to understand the type of results that each technique delivers based on its computational algorithm [13] so that an appropriate one can be selected for a specific purpose.

In this paper, the Coronavirus data that consist of the reported number of cases and deaths for 3,143 counties in the United States are segmented into clusters so that each cluster represents a geographical area where the viruses are spreading. Each data point $d_n(t)$ for $1 \leq n \leq 3143$, consists of a quartet $(\kappa_n(t), \delta_n(t), \lambda_n, \nu_n)$ where $\kappa_n(t)$ is the number of confirm cases at time t, $\delta_n(t)$ is the number of confirmed deaths at time t, λ_n the latitude and ν_n the longitude in the geographical coordinates [14] of the center of the county indexed by n. The integer index n used in this paper is an index assigned to a county in the Federal Information Processing Standards system [15]. Due to the ongoing development of the Coronavirus pandemic, no known statistical properties of the data are available, and therefore the K-Means clustering technique [16] is selected to segment the data.

The K-Means technique requires the use of distance function [17] to measure similarities between data points and clusters so that they can be assigned to appropriate cluster. Due to the requirement of the K-Means technique that the number of clusters must be known prior to applying the algorithm to divide the data set into clusters, an algorithm similar to that of the model-based technique [18] is developed to approximate the number of clusters at the beginning before applying the K-Means technique. In this development, it is assumed that an epicenter of the pandemic crisis has a bell shape with its peak at the center and its surroundings having smaller values the further they are from that center. This bell shape can be formulated mathematically, and the semblance value is calculated by correlating the bell shape function at various locations with the data set. The locations of the local maxima are identified as the epicenters.

The K-Means clustering technique with this newly developed initial algorithm for identifying the number of clusters is used to segment the Coronavirus data into clusters so that each cluster can be seen as an epicenter where the number of cases and deaths for the proximity can be

TABLE I. Example of Daily Report of Known Cases of Coronavirus in the United States*.

	Daily Number of Confirmed Cases							
County	Oct. 11,2020	Oct. 12,2020	Oct. 13,2020	Oct. 14,2020	Oct. 15, 2020	Oct. 16, 2020	Oct. 17, 2020	Total Last 7 Days
Cook County, IL	1005	1048	1159	847	1462	1610	1575	8706
Los Angeles County, CA	970	847	768	1266	1167	1039	914	6971
El Paso County, TX	546	0	885	503	0	1591	582	4107
Miami-Dade County, FL	1006	279	440	434	538	530	554	3781
Salt Lake County, UT	0	925	342	548	570	706	582	3673
Harris County, TX	450	535	302	417	743	354	594	3395
Dallas County, TX	460	418	446	606	454	537	462	3383
Maricopa County, AZ	431	296	436	541	601	460	590	3355
Tarrant County, TX	335	683	289	454	436	522	494	3213
Milwaukee County, WI	342	235	555	511	538	400	0	2581

*Source: https://usafacts.org

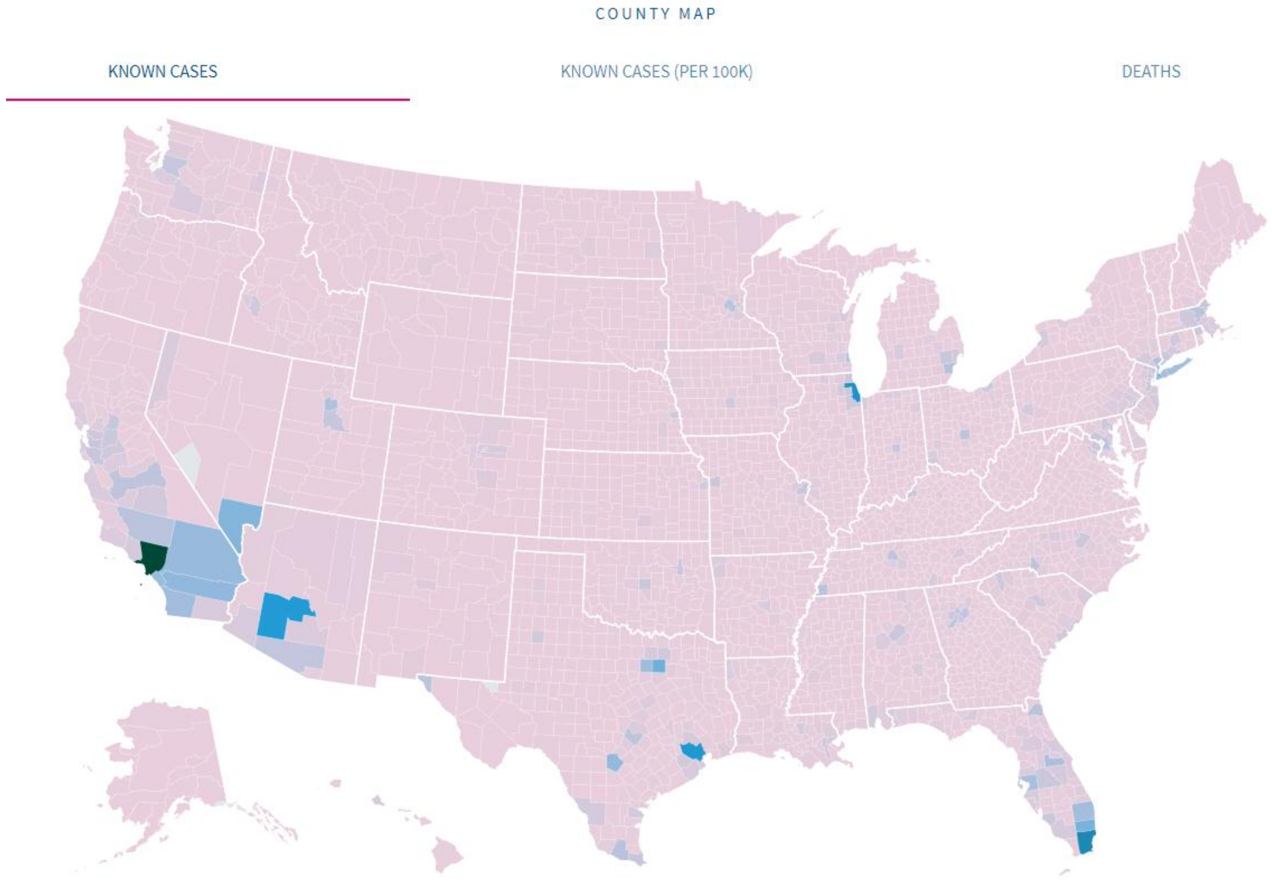
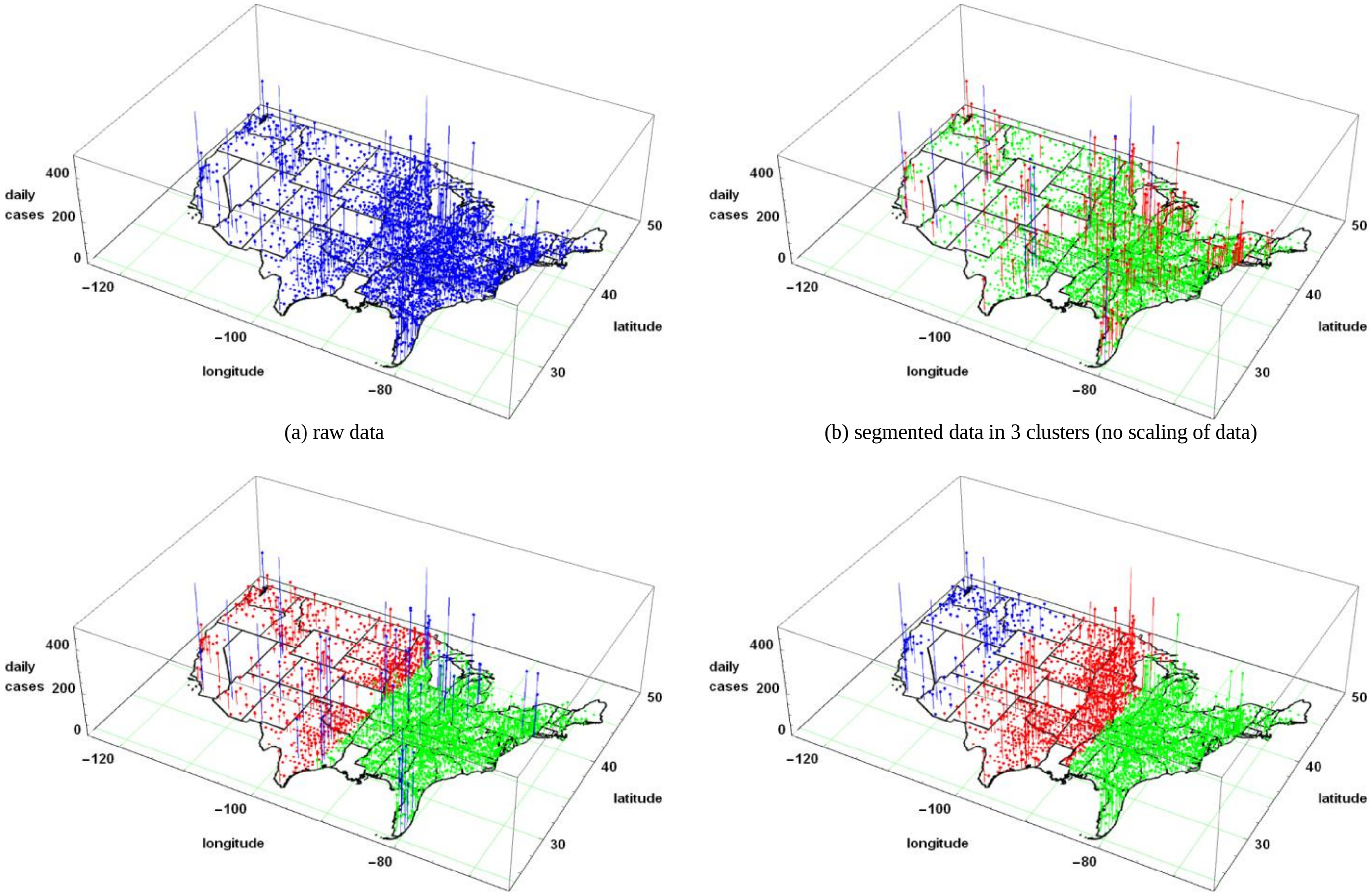


Fig. 1. Example of Graphical Representation of a Daily Report (Oct. 18, 2020) of Coronavirus in the United States (Source: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/).

identified. Numerical results are included to visually demonstrate the workability and applicability of the concept.

II. Background

Coronavirus data are recorded daily and newspapers have been posting them for the public to view. Table I shows the typical format that data are posted to the public summarily showing the total number of confirmed cases in the last seven days and the daily number of newly confirmed cases from each of the last seven days. Since there are too many counties to report, the data are often truncated to a handful of counties with the highest number of confirmed cases. Thus, attention is often directed to the counties with the most severe effect of the Coronavirus pandemic. In this aspect, it is possible for humans to only see the individual counties that have already reached their saturation points while unintentionally ignoring the areas that are on the rise and spreading out and consequently require utmost attention. Even with the help of graphics such as the map shown in Fig.1, the human attention often can only direct at one county at a time without the capability to effectively



(a) raw data

(b) segmented data in 3 clusters (no scaling of data)

(c) segmented data in 3 clusters (20% scaling on z-axis of case counts)

(b) segmented data in 3 clusters (10% scaling on z-axis of case counts)

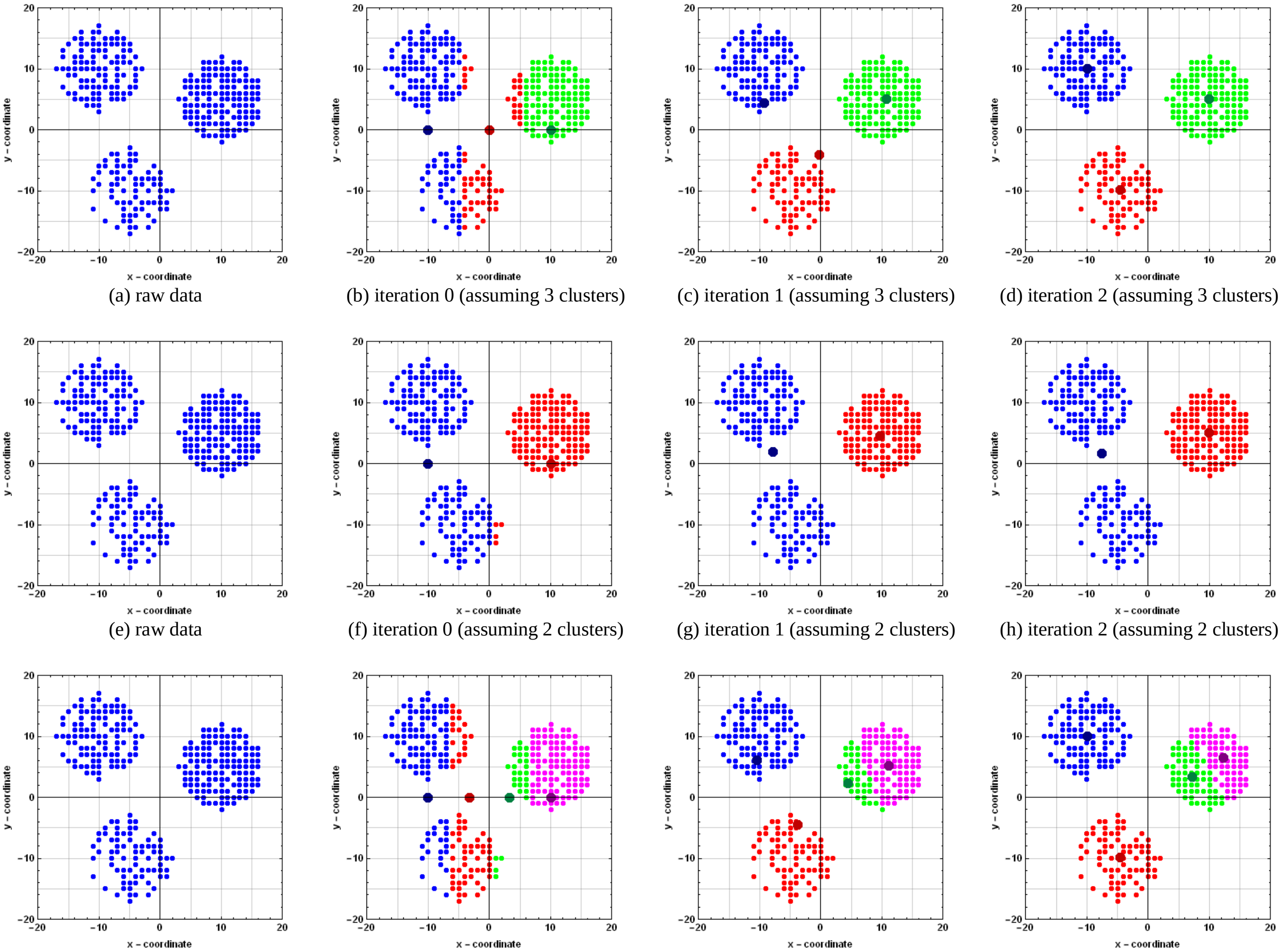
Fig. 2. Segmentation of Coronaviurs data in the US (Oct. 17, 2020) with K-Means clustering technique, with the assumption of there exist only 3 clusters.

connect it to the other counties in the same proximity that have similar numbers of cases or deaths to form a model of disease spreading out for further tracking and studying.

In order to identify the areas that are on the rise, various counties that are situated within the proximity of each other must be grouped together according to their similarities in the data values so that they can be analyzed at the cluster level instead of at the individual county level. In this endeavor, it is proposed to use the first stage of segmentation of the data mining process so that humans do not have to be involved to aggravate the results with unintended mistakes when dealing with a large quantity of data in a repetitive manner.

While there are four basic families of clustering techniques: K-means, hierarchical, fuzzy C-means, and model-based; these techniques can offer different results when applied to the same set of data. The K-means clustering technique [19] is the simplest and most computationally efficient because it only examines the similarities between a data point to each of the centers of the clusters and assigns it to the cluster that shows the most similarity. The fuzzy C-means clustering technique [20] is a variation of the K-means technique, with fuzzy membership function being used to deal with data that are approximately or numerically equidistant to many clusters. The hierarchical clustering technique [21] is computationally intense because examines every possible pair of data points in each iteration to group them in a growing manner. The model-based clustering technique [22] assumes that data follow certain statistical properties and assigns them to a cluster that increases the likelihood of that statistical assumption.

For analyzing the areas that are on the rise in the case counts and death counts of the Coronavirus infection, the daily reports are segmented into clusters based on the geographical coordinates, the daily number of confirmed cases, and the daily number of confirmed deaths for each county. Since there exists no knowledge of the statistical properties of these data, the K-means clustering technique is selected. However, the K-means clustering technique (as well as the other three clustering techniques) requires the knowledge of the number of clusters in advance, making it hard to extract any meaningful information from the resulting clusters. Fig. 2 shows the resulting segmentation of the Coronavirus data collected on October 17, 2020 into three arbitrary clusters. In Fig. 2(a), the complete data set is plotted across the map of the United States, with the horizontal axes showing longitude and latitude coordinates, and the vertical axis showing the number of case counts. Fig. 2(b) shows the three clusters plotted in three distinct colors red, green, and blue. Here, the magnitude



(a) raw data (b) iteration 0 (assuming 3 clusters) (c) iteration 1 (assuming 3 clusters) (d) iteration 2 (assuming 3 clusters)

(e) raw data (f) iteration 0 (assuming 2 clusters) (g) iteration 1 (assuming 2 clusters) (h) iteration 2 (assuming 2 clusters)

(i) raw data (j) iteration 0 (assuming 4 clusters) (k) iteration 1 (assuming 4 clusters) (l) iteration 7 (assuming 4 clusters)

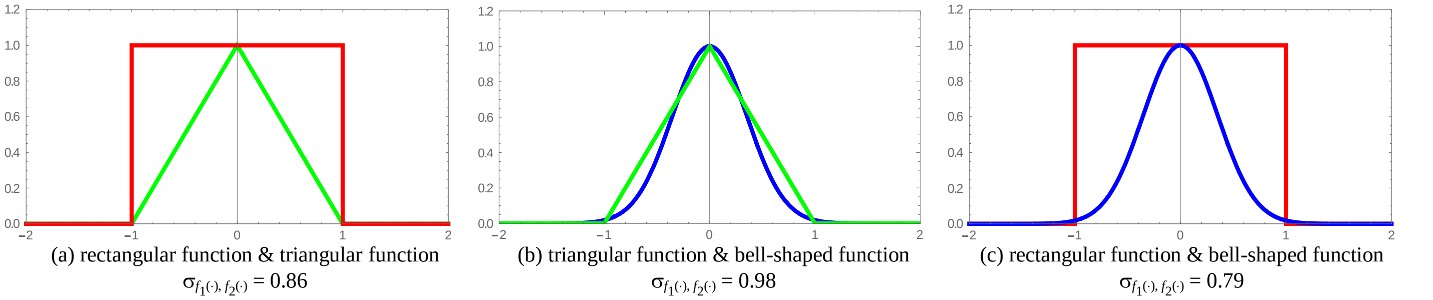
Fig. 3. Examples of K-Means clustering technique with different initial conditions of number of clusters for the same data set.

of the data on the vertical axis dominates the magnitude of the data on the other two axes, the clusters show separation along the vertical z-axis. Fig. 2(c) shows the same segmentation with K-means clustering technique as in Fig. 2(b), except with the data on the z-axis being scaled to 20%. The result still shows a smaller dominance of the vertical z-axis but with better defined geographical areas for the clusters. Fig. 2(d) shows the same results as Fig. 2(c), but with the data on the z-axis being scaled to 10%. Here the result show a balance of clusters spreading across the geographical coordinates.

As illustrated in Figure 2, knowledge of the number of clusters must be available before applying the clustering technique to a data set. Furthermore, the data must be scaled appropriately so that meaningful information can be extracted from the resulting clusters. These requirements defeat the purpose of using data mining to discover information. In the case of the Coronavirus data in the US, the important information needed is the clusters representing an area surrounding an epicenter of the pandemic crisis so that further studies such as modelling of the spreading can be conducted. Thus, the methodology in the next section is developed in order to address the need to get around the requirements of the K-means clustering technique with modification while still allowing the clustering technique to be used in the absence of knowledge of the statistical properties of the data set.

III. Methodology

K-means clustering technique is a method of dividing a data set into several subsets according to the similarity in the values of the elements in the data set (Fig. 3). The similarity is measured through a distance function applied to two elements: the smaller the distance between them the more similar they are assumed to be. One advantage of K-means clustering technique over other techniques is that it does not require any prior knowledge of the statistical properties of a data set. A disadvantage of all clustering techniques is that they all require knowledge of how many subsets there are in a data set at the beginning. In this section, the algorithm for the K-means

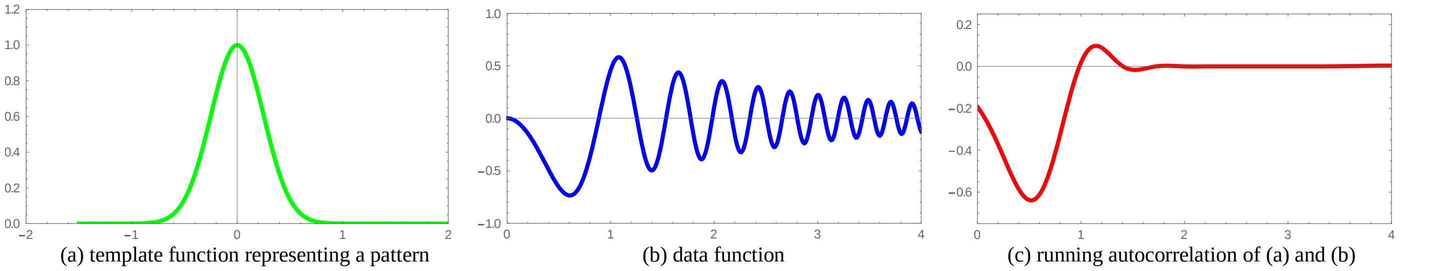


(a) rectangular function & triangular function $\sigma_{f_1(\cdot),f_2(\cdot)} = 0.86$

(b) triangular function & bell-shaped function $\sigma_{f_1(\cdot),f_2(\cdot)} = 0.98$

(c) rectangular function & bell-shaped function $\sigma_{f_1(\cdot),f_2(\cdot)} = 0.79$

Fig. 4. Examples of semblance test with normalized correlation (the higher the value of σ, the more similar are the two functions).



(a) template function representing a pattern

(b) data function

(c) running autocorrelation of (a) and (b)

Fig. 5. Examples of running autocorrelation, with the maximum in (c) representing the location where the pattern in (a) appears in (b).

clustering technique is described, then the formulation of the semblance method commonly used to find a pattern that occurs in a data set is presented, and a mathematical model of a pandemic in development is presented so that the semblance method can be applied to a data set to find the approximate locations of the epicenters as an initial condition for the K-means clustering algorithm to segment the data set.

A. Algorithm for K-Means Clustering Technique

Consider the Coronavirus data set $D(t) = \{ d_n(t) = (\eta_n(t), \lambda_n, \nu_n) \mid 1 \le n \le 3143 \}$, where $\eta_n(t)$ is the number of confirmed cases, λ_n and ν_n the geographical longitude and latitude of a county indexed by an integer n in the Federal Information Processing Standards system. The K-Means clustering technique used for segmenting the data set D consists of the following steps in its algorithm:

(i) Initialize the algorithm with an integer constant K, the assumed number of clusters $\kappa_1, \kappa_2, \ldots, \kappa_K$,

(ii) Initialize the K centers $c_1, c_2, \ldots, c_K$ for the K clusters $\kappa_1, \kappa_2, \ldots, \kappa_K$ with arbitrary data constants

(iii) For each data point d_n, calculate the distances $\Delta(d_n, c_k)$ for $1 \le k \le K$ from d_n to each of the K centers $c_1, c_2, \ldots, c_K$

(iv) Select the index k^* for $1 \le k^* \le K$ for the minimum distance $\Delta(d_n, c_{k^*}) \le \Delta(d_n, c_k)$ among the distance values $\Delta(d_n, c_k)$ for $1 \le k \le K$ calculated in Step (iii)

(v) Assign the data d_n to the cluster κ_{k^*}

(vi) Recalculate the K centers $c_1, c_2, \ldots, c_K$ after all data points d_n for $1 \le n \le 3143$ have been assigned to the K clusters $\kappa_1, \kappa_2, \ldots, \kappa_K$

(vii) Repeat Steps (iii) through (vi) until the changes to the centers $c_1, c_2, \ldots, c_K$ are numerically negligible

The distance function $\Delta(\cdot,\cdot)$ is a class of analytical functions that satisfy the following properties:

(i) $\Delta(p_1, p_2) \ge 0 \ \forall \ p_1, p_2$,

(ii) $\Delta(p_1, p_2) = \Delta(p_2, p_1)$,

(iii) $\Delta(p_1, p_2) = 0 \Leftrightarrow p_1 = p_2$,

(iv) $\Delta(p_1, p_2) + \Delta(p_2, p_3) \ge \Delta(p_1, p_3)$.

In the context of this paper, the Euclidean distance function [17] is used due to its simplicity and common use in our 3-dimensional world:

$$\Delta(p_1, p_2) = \sqrt{(p_{1,1} - p_{2,1})^2 + \ldots + (p_{1,n} - p_{2,n})^2}\ , \qquad (1)$$

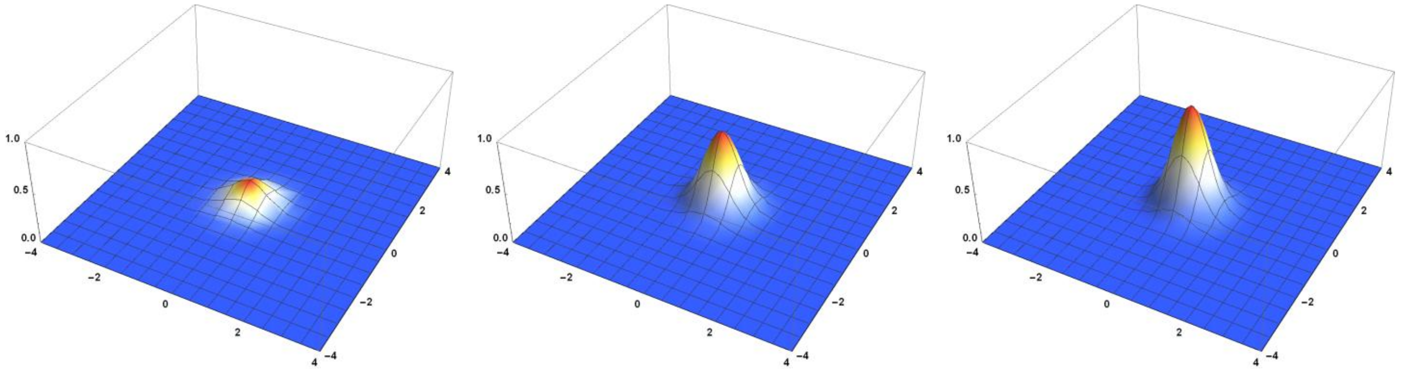
where $p_1 = [p_{1,1}\ \ p_{1,2}\ \ldots\ p_{1,n}]^T$ and $p_2 = [p_{2,1}\ \ p_{2,2}\ \ldots\ p_{2,n}]^T$.

Fig. 3 shows an example of the K-means clustering algorithm applied to a small set of data for visual demonstration of how the algorithm works. While there are many iterations involved in the algorithm, it is considered computationally efficient when compared to the other three clustering techniques (fuzzy C-means, hierarchical, and model based). The first two steps in the algorithm constitute an initialization of the algorithm that will affect how fast the algorithm converges to the final solution. Thus, it is important to know the "correct" number of clusters at the beginning in Step (i), and to "accurately" guess the initial centers of these clusters in Step (ii).

B. Semblance Method

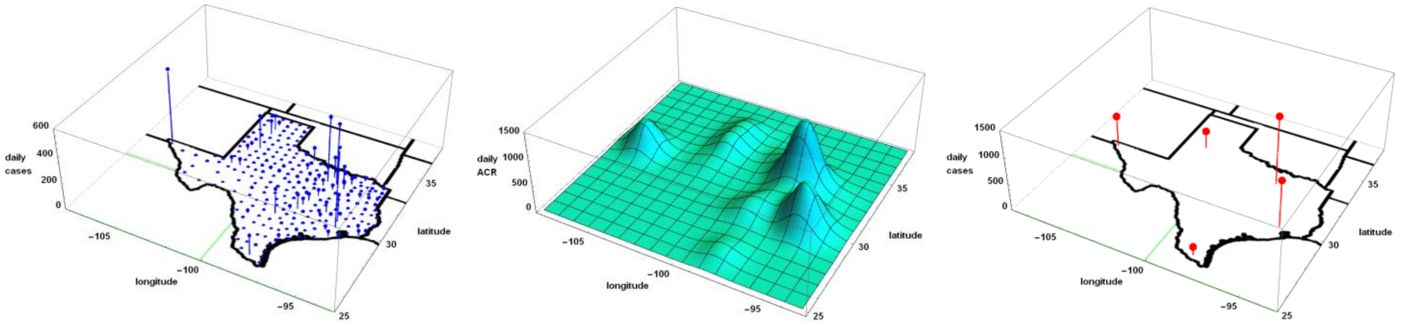
Semblance method [23] is the method of identifying similarity between a known pattern and a sequence of data. Typically, the normalized correlation formula is used to determine if two functions (or two data series) are similar:

$$\sigma_{f_1(\cdot),f_2(\cdot)} = \frac{\int \ldots \int f_1(t_1,\ldots,t_N) f_2(t_1,\ldots,t_N)\, dt_1 \ldots dt_N}{\sqrt{\int \ldots \int f_1^2(t_1,\ldots,t_N)\, dt_1 \ldots dt_N}\ \sqrt{\int \ldots \int f_2^2(t_1,\ldots,t_N)\, dt_1 \ldots dt_N}}\ , \qquad (2)$$



(a) $t = t_0 + \Delta t$ (b) $t = t_0 + 3\Delta t$ (c) $t = t_0 + 4\Delta t$

Fig. 6. Example of the modeling of an area of outbreak as an analytical function $f(x, y, t) = \alpha(t) \cdot \exp(-2(x^2 + y^2)/\sigma^2)$



(a) Texas data (b) running autocorrelation data (c) epicenters

Fig. 7. Example of using the outbreak model and the running autocorrelation to approximately identify the epicenters of Coronavirus pandemic

where the scalar value $\sigma(f_1(\cdot), f_2(\cdot))$ has the range from –1 to 1. A value of 0 means no resemblance between the two functions, and a value of 1 means an exact duplication $f_1(\cdot) = f_2(\cdot)$, and a value of –1 means the negative duplication $f_1(\cdot) = -f_2(\cdot)$. Fig. 4 shows examples of the use of normalized correlation to determine the similarity between two functions.

To use the normalized correlation formula (2) for approximately determining the epicenters for initialize the K-means clustering technique, two elements will be needed: (i) an approximate model of the area surrounding an epicenter of a pandemic, and (ii) an algorithm to determine the local maxima of a running autocorrelation [24] formula derived from (2):

$$\sigma_{f_1(\cdot), f_2(\cdot)}(\tau_1, \tau_1, \ldots, \tau_N) =$$

$$\frac{\int \ldots \int f_1(t_1 - \tau_1, \ldots, t_N - \tau_N) f_2(t_1, \ldots, t_N) dt_1 \ldots dt_N}{\sqrt{\int \ldots \int f_1^2(t_1, \ldots, t_N) dt_1 \ldots dt_N} \sqrt{\int \ldots \int f_2^2(t_1, \ldots, t_N) dt_1 \ldots dt_N}}, \quad (3)$$

where $(\tau_1, \tau_1, \ldots, \tau_N)$ is location that $f_1(\cdot)$ is shifted to to determine if there is a resemblance with $f_2(\cdot)$ at that location. Fig. 5 shows the running autocorrelation between two functions where the location that the maximum value occurs represent the location that the similarity is identified.

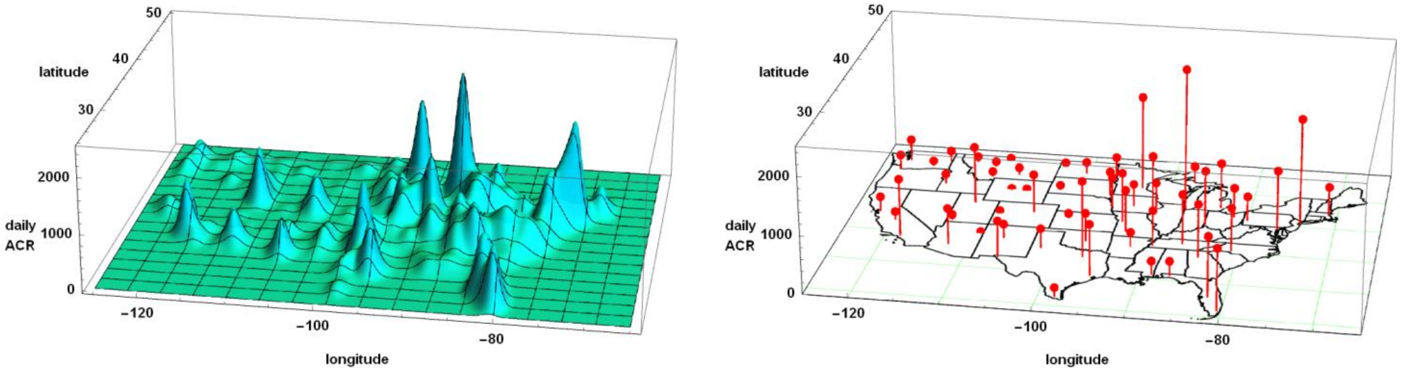
C. *Algorithm of Determining the Number of Clusters*

An area containing an outbreak of the Coronavirus (or any pandemic disease) will start at an epicenter where a number of people got infected. This number starts to increase with time while people in the surrounding area start to get infected. Fig. 6 shows a typical mathematical model $f(x, y, t)$ that fits this description:

$$f(x, y, t) = \alpha(t) \cdot \exp\left(-\frac{2}{\sigma^2}(x^2 + y^2)\right), \quad (4)$$

where $\alpha(t)$ is the growth function, and $\exp(-2(x^2 + y^2)/\sigma^2)$ is the bell-shaped model with an approximate radius σ. Here in this paper, the study is conducted for a snapshot in time and therefore the growth function $\alpha(t)$ is not necessary and will be set to the value of 1. This growth function will be discussed in the scope of the future work where daily snapshot data are put together as a function of time.

The pandemic model in (4) will be shifted and correlated with the pandemic data in a running autocorrelation formula in (3) and the results will show a smooth function with various local maxima representing the epicenters where the case counts reach the highest in comparison to the case counts in their immediate surroundings. Fig. 7 shows an example of this running autocorrelation for the pandemic model in (4) with the pandemic data in Texas recorded on October 17, 2020. In Fig. 7(a), the pandemic data are plotted for various counties. In Fig. 7(b), the running autocorrelation is shown with a few local maxima. In Fig. 7(c), the local maxima are located and plotted on the map of Texas, showing epicenters at El Paso (the furthest western point), Lubbock (the panhandle), Dallas-Fort Worth (the northern point toward the eastern boundary), Houston (the southeastern coastal point), and Brownsville-McAllen (the southernmost tip). These areas are the hotspots under serious watch by health officials, thus confirming the



(a) running autocorrelation

(b) local maxima (epicenters)

Fig. 8. Using autocorrelation in (a) between pandemic model and Coronavirus data in the US (Oct. 17, 2020) to identify epicenters in (b) of the pandemic crisis.

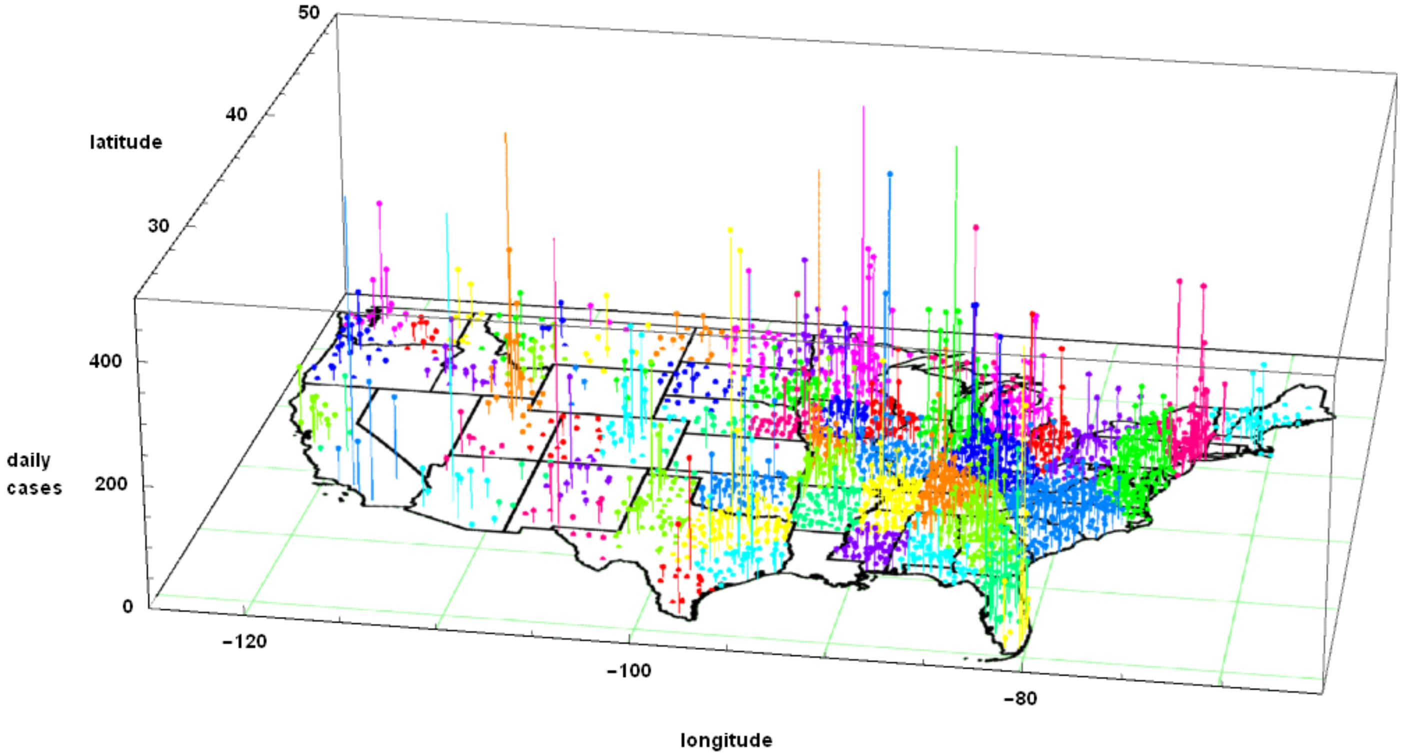


Fig. 9. Using K-means clustering technique with identified epicenters shown in Fig. 8(b) to segment the Coronavirus data in the US (Oct. 17, 2020).

workability of the running autocorrelation. After calculating the running autocorrelation, it is important to identify these local maximum points so that they can be used as the epicenters for the initial condition of the K-means clustering technique.

Given the resulting matrix of running autocorrelation $\Phi(n, m)$ for $0 \le n < N$ and $0 \le m < M$, the local maximum points of $\Phi(n, m)$ are identified in the following algorithm:

(i) Initialize the set $\mu = \{ \varnothing \}$ as empty set

(ii) Start at index $(n, m) = (0, 0)$ and increase n and m in a double nested loop structure until $(N-1, M-1)$

(iii) For each value $\Phi(n, m)$, examine its surrounding points $\Phi(n-1, m-1)$, $\Phi(n, m-1)$, $\Phi(n+1, m-1)$, $\Phi(n-1, m)$, $\Phi(n+1, m)$, $\Phi(n+1, m-1)$, $\Phi(n+1, m)$, and $\Phi(n+1, m+1)$. If $\Phi(n, m)$ is greater that all these surrounding points, add the data point (n, m) to the set μ.

(iv) At the end of the nested loop structure, the set μ will contain the local maximum points of the running autocorrelation $\Phi(n, m)$.

It is common to consider the value $\Phi(n, m)$ in Step (iii) when it has the value between 10% and 100% of the global maximum of Φ. In this case, the global maximum is searched first, and will be used to calculate the thresholds for filtering out data of small values.

IV. Numerical Results

The Coronavirus data in the US for October 17, 2020, summarily shown in Fig. 2(a), is segmented using the steps outlined in the previous section. The intermediate results are

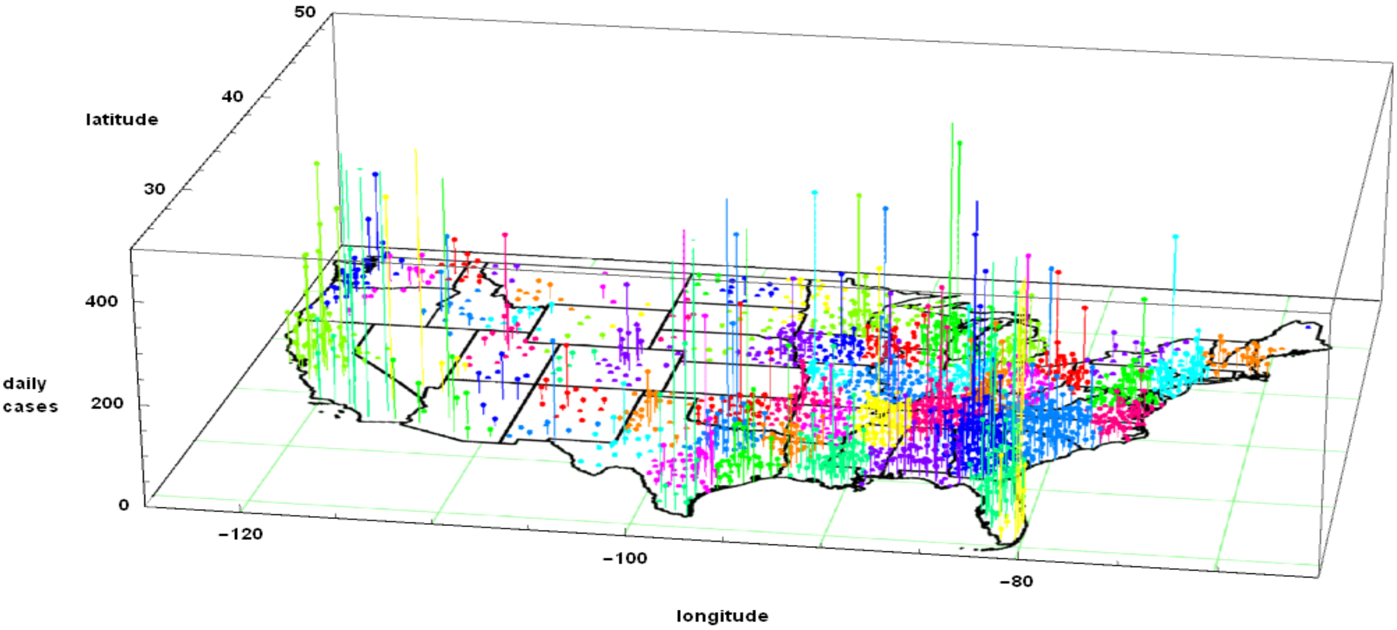


Fig. 10. Using K-means clustering technique with epicenters identified by running autocorrelation to segment the Coronavirus data in the US (Jul. 23, 2020).

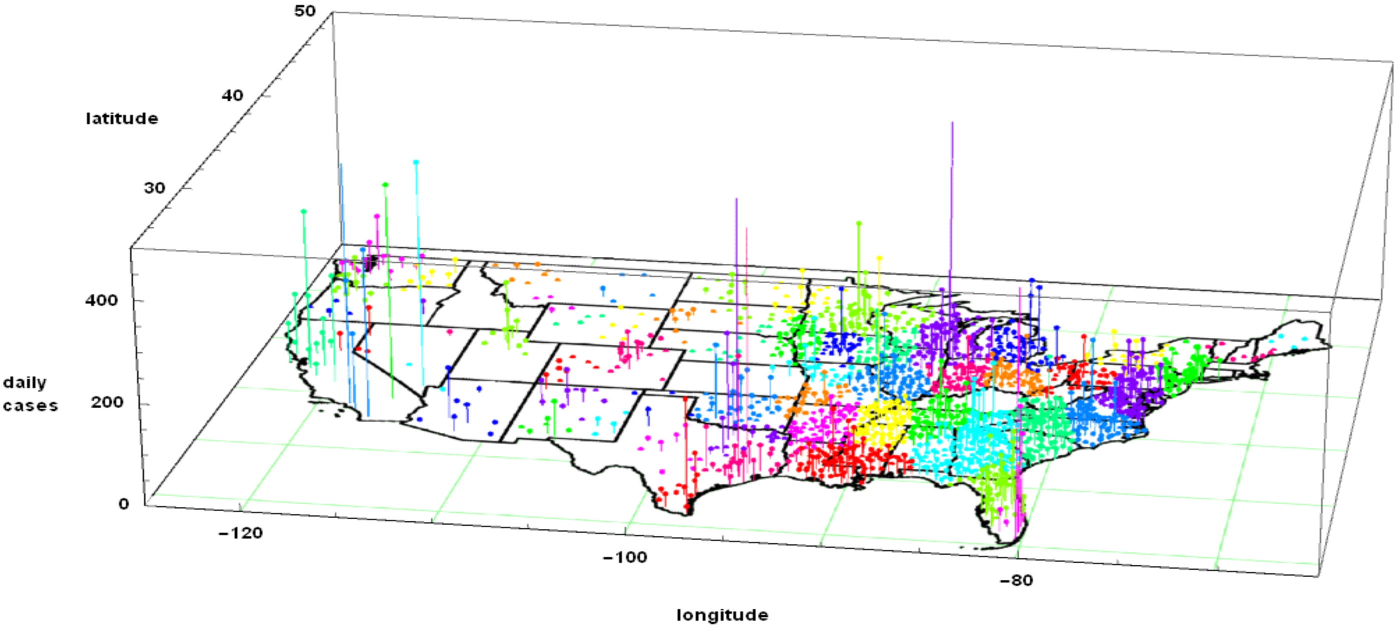


Fig. 11. Using K-means clustering technique with epicenters identified by running autocorrelation to segment the Coronavirus data in the US (Aug. 23, 2020).

shown in Fig. 8, and the final results are shown in Fig. 9. Due to a large number of epicenters (around 60) detected in the step of running autocorrelation and shown in Fig. 8, it is not efficient to manually assign colors to each cluster in Fig. 9. Therefore, an automatic color assignment scheme is used [25] to arbitrary assign a contrasting color to the cluster as it appear on the list. Here, the results can visually be verified with the region of Texas where the epicenters were shown previously in Fig. 7. The five clusters associated with the five epicenters (El Paso, Lubbock, Dallas-Fort Worth, Houston, and Brownsville) are seen correctly identified.

Figs. 10-13 show the segmentation for the Coronavirus data for the four most recent months of July 23, August 23, September 23, and October 23 of 2020. Here the snapshot data are shown once for each of the months so that the difference can be examined visually. In reality, the data can be automatically prepare on the daily basis to make a video sequence so that the gradual change can be observed on the slow motion. On the west coast, the situation was severe in July, but improved steadily in August and September, and stabilized in October. This improvement was probably due to the precautious measure imposed by the government. On the East Coast and in the Northeast industrial areas, the situation was severe in July, improved in August and September, but

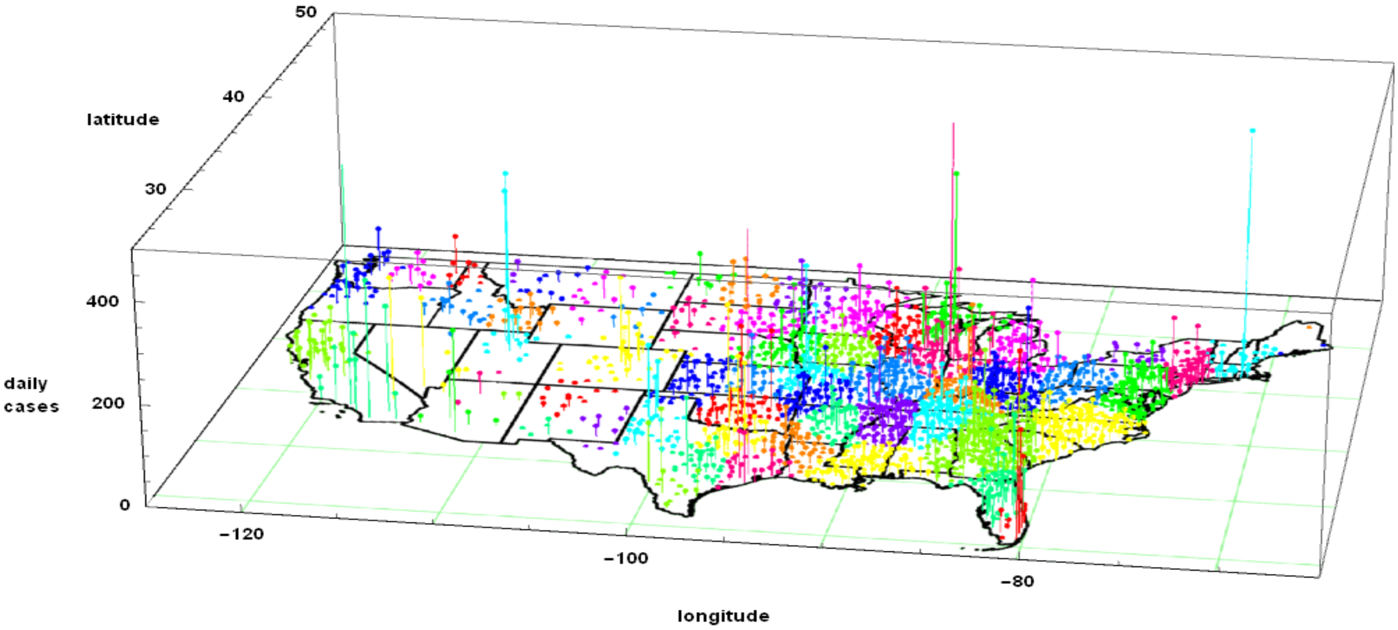


Fig. 12. Using K-means clustering technique with epicenters identified by running autocorrelation to segment the Coronavirus data in the US (Sep. 23, 2020).

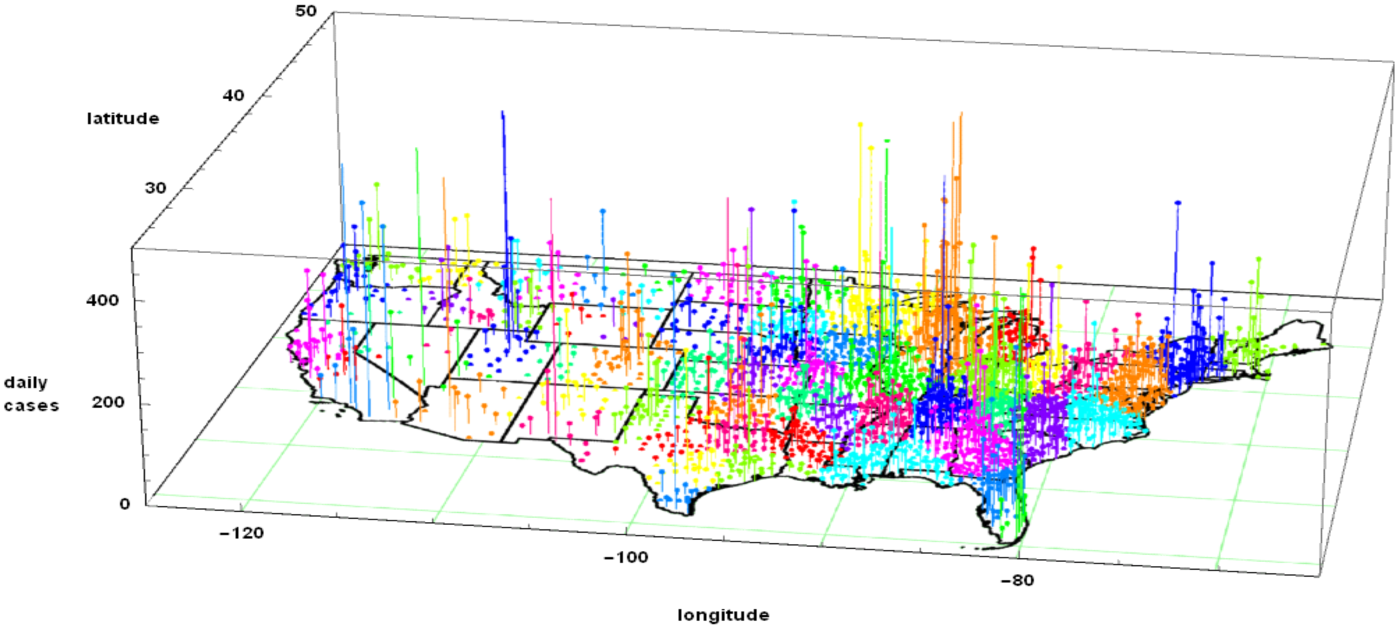


Fig. 13. Using K-means clustering technique with epicenters identified by running autocorrelation to segment the Coronavirus data in the US (Oct. 23, 2020).

reverted back to severe in October. This recent increase in cases is probably due to the reopening of schools and businesses in an attempt to restart the economy that has been paralyzed during earlier months. Elsewhere in the South and Midwest, new spikes are observed in Texas, Colorado, Utah, etc. during October. Again, these new spikes are probably due to the attempt to restart the economy and the reopening of schools.

V. Discussion and Future Direction

It has been shown that the segmentation of the Coronavirus data in the US can be done individually on a daily basis. The results can be put together in a time sequence and shown as a slow motion movie for visual observation. However, the segmentation technique can easily be extended from three-dimensional data to four-dimensional data that include time as an additional dimension. The K-means clustering technique is not limited to any particular dimension and can be arbitrarily extended without any restriction. However, the initial condition of the presumed number of clusters for the K-means clustering technique is still required to be established.

The immediate future work for this project is to find an approximate model of the pandemic expansion both in time and in spatial coordinates so that it can be correlated with the four dimensional data in order to identify the epicenters of the Coronavirus crisis. These epicenters will be used as initial conditions for the K-means clustering technique to find the clusters of surrounding areas with similar number of confirmed cases. Furthermore, the number of confirmed deaths can also be included as an additional dimension to complete the pandemic model.

VI. Conclusion

The Coronavirus data in the US have been segmented by the K-means clustering technique on the daily basis for visual analysis. In this segmentation, the initial condition of the number of epicenters has been automated with the running autocorrelation formula where a spiking model is used against the data. Numerical results have been shown in three-dimensional visualization, with colors being used to differentiate different clusters, for human experts to analyze. Future work of expanding the data to include time dimension and death counts was outlined as the continuation of the project.

Acknowledgment

Data concerning the reported cases of Coronavirus infection and the resulting deaths used in this study were provided by the US Government Center for Disease Control and Prevention (http://www.cdc.gov), and the not-for-profit, nonpartisan civic initiative USA Facts organization (http://www.usafacts.org). Data concerning the geographical attributes of the 3143 counties in the US used in this study were provided by the information website Wikipedia (https://en.wikipedia.org/wiki/User:Michael_J/County_table).

References

[1] P. Yang, & X. Wang. “COVID-19: A New Challenge for Human Beings,” *Cellular & Molecular Immunology*, 17, pp 555-557, 2020.

[2] Centers for Disease Control and Prevention. COVIDView – A Weekly Surveillance Summary of U.S. COVID-19 Activities, 2020. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/index.html

[3] E. Toch, B. Lerner, & E. Ben-Zion. “Analyzing large-scale human mobility data: a survey of machine learning methods and applications,” *Knowledge & Information Systems*, 58, pp 501–523, 2019.

[4] J. B. Schmitt, C. A. Debbelt, & F. M. Schneider, F. M. “Too much information? Predictors of information overload in the context of online news exposure,” *Journal of Information, Communication & Society*, 21(8), pp 1151-1167, 2017.

[5] S. Khanam, J. Siddiqui, & F. Talib. (2016). “Role of Information Technology in Total Quality Management: A Literature Review,” *International Journal of Advanced Research in Computer Engineering and Technology*, 2(8), pp 2433-2445, 2016.

[6] M. A. Virtanen, E. Haavisto, E. Liikanen, & M. Kääriäinen. “Ubiquitous learning environments in higher education: A scoping literature review,” *Education and Information Technologies*, 23, pp 985–998, 2018.

[7] Y. C. Ko, & H. Fujita. “An evidential analytics for buried information in big data samples: Case study of semiconductor manufacturing,” *Information Sciences*, 486, pp 190-203, 2019.

[8] S. Pal, P. K. D. Pramanik, T. Majumdar, & P. Choudhury. “A semi-automatic metadata extraction model and method for video-based e-learning contents,” *Education and Information Technologies*, 24, pp 3243–3268, 2019.

[9] P. N. Tan, M. Steinbach, A. Karpatne, & V. Kumar. *Introduction to Data Mining*. New York, NY: Pearson Publishing, 2018.

[10] G. Persad. “Beyond Administrative Tunnel Vision: Widening the Lens of Costs and Benefits,” *Georgetown Journal of Law & Public Policy*, 15, pp 941-957, 2017.

[11] M. J. Zaki, & W. Meira. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. New York, NY: Cambridge University Press, 2020.

[12] A. Langley, & H. Tsoukas. *The SAGE Handbook of Process Organization Studies*. Thousand Oaks, CA: SAGE Publications, 2017.

[13] R. Xu, & D. Wunsch. *Clustering*. Hoboken, NJ: Wiley - IEEE Press, 2008.

[14] Y. Zhong, J. Liu, & B. Hu. “Mathematical Transformation of Latitude and Longitude Network on Several World Map Projections,” *Geomatics Science and Technology*, 7(2), pp 117-123, 2019.

[15] NIST. *Federal Information Processing Standards Publications*. Washington, DC: National Institute of Standards and Technology, 2017.

[16] J. Wang, J. “Generalized 2-D Principal Component Analysis by Lp-Norm for Image Analysis,” *IEEE Transactions on Cybernetics*, 46(3), pp 792-803, 2016.

[17] L. Debnath, & P. Mikusinski. *Introduction to Hilbert Spaces with Applications. Burlington*, MA: Academic Press, 2005.

[18] C. Bouveyron, & C. Brunet-Saumard. “Model-based clustering of high-dimensional data: A review,” *Computational Statistics & Data Analysis*, 71, pp 52-78, 2014.

[19] L. Morissette, & S. Chartier. “The k-means clustering technique: General considerations and implementation in Mathematica,” *Tutorials in Quantitative Methods for Psychology*, 9(1), pp 15-24, 2013.

[20] J. C. Bezdek, R. Ehrlich, & W. Full. “FCM: The fuzzy c-means clustering algorithm,” *Computers & Geosciences*, 10(2), pp 191-203, 1994.

[21] F. Murtagh. “A survey of recent advances in hierarchical clustering algorithms,” *The Computer Journal*, 26(4), pp 354-359, 1983.

[22] S. Zhong, & J. Ghosh. “A Unified Framework for Model-based Clustering,” *Journal of Machine Learning Research*, 4(11), pp 1001-1037, 2003.

[23] G. R. J. Cooper, & D. R. Cowan. “Comparing time series using wavelet-based semblance analysis,” *Computers & Geosciences*, 34(2), pp 95-102, 2008.

[24] D. A. Griffith. *Spatial Autocorrelation and Spatial Filtering*. New York, NY: Springer, 2003.

[25] J. T. Pham. “Segmentation of Electron Microscopic Images with Artificial Gravitational Field for Enhancement of Visibility,” *IOSR Journal of Computer Engineering*, 22(5), pp 12-31, 2020.