Scientific Research Publishing

Journal of Computer and Communications, 2025, 13(9), 1-12
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Application of the K-Means Algorithm in the Study of Influenza Transmission Patterns

Shuyuan Ye

Guangdong Experimental High School, Guangzhou, China
Email: 13910921634@139.com

How to cite this paper: Ye, S.Y. (2025) Application of the K-Means Algorithm in the Study of Influenza Transmission Patterns. *Journal of Computer and Communications*, **13**, 1-12.
https://doi.org/10.4236/jcc.2025.139001

Received: August 4, 2025
Accepted: August 31, 2025
Published: September 3, 2025

Abstract

As a highly contagious respiratory disease, influenza exhibits significant spatiotemporal fluctuations in incidence, posing a persistent threat to public health and placing considerable strain on healthcare resource allocation and emergency response systems. Accurately grasping its epidemic characteristics is crucial for improving prevention and control efficiency. This study selects 29 cities as the research subjects and employs K-means clustering to classify them based on three core indicators: administrative level, GDP, and influenza incidence rate. The optimal number of clusters is determined using the elbow method, and MATLAB is used for data processing and model computation. The findings reveal that cities with higher GDP and administrative level tend to have lower incidence rates, likely due to more abundant medical resources and robust prevention systems. In contrast, cities with lower GDP and administrative levels generally exhibit higher incidence rates due to limited resource allocation. The results provide a scientific basis for developing differentiated influenza prevention strategies and optimizing the allocation of public health resources.

Keywords

K-Means Algorithm, Influenza Transmission, Cluster Analysis, Urban Characteristics

1. Introduction

As a seasonal epidemic, influenza imposes a substantial burden on global public health systems each year, particularly in densely populated urban regions. Its transmission is closely associated with factors such as population mobility, climatic conditions, and socioeconomic dynamics. Therefore, a rigorous and scientific analysis of the spatiotemporal distribution characteristics of influenza incidence is essential for enhancing the effectiveness of disease prevention and control strategies.

In recent years, with the rapid development of data mining and machine learning technologies, the K-means algorithm—an unsupervised learning method—has demonstrated considerable utility in multiple domains by categorizing data into clusters based on similarity. For instance, Han Xiaocui *et al.* [1] employed the algorithm for anomaly detection in human resources management, while Weng Ziyun [2] applied it to fault detection in DC power grid converters. In the field of public health, this algorithm likewise assists researchers in identifying regions with high incidence rates, analyzing potential risk factors, and optimizing the allocation of public health resources.

In the context of influenza incidence studies, scholars have attempted to explore its spatial distribution patterns using clustering methods. Bao Weina [3] proposed the application of the K-means++ algorithm to perform sequence clustering and labeling, offering new insights for the characterization of influenza viruses. Xia Hu *et al.* [4] employed the K-means clustering model to analyze the spatial aggregation of cases and explored the integration of artificial intelligence into hospital-based infectious disease early warning systems. Their findings demonstrated that combining clustering with LSTM prediction models can effectively forecast influenza case numbers, providing critical support for the formulation of prevention strategies.

Moreover, the fundamental principles and optimization of the K-means algorithm have been extensively studied. Liu Fugang [5] pointed out that the algorithm enables data mining through group analysis by ensuring intra-cluster similarity while highlighting inter-cluster differences. To address the sensitivity of traditional K-means to initial cluster centers, Zhou Xiaodong *et al.* [6] proposed a geometry-based optimization approach to enhance clustering stability.

Despite its demonstrated utility in analyzing influenza incidence, the K-means algorithm is not without limitations. For instance, it is sensitive to initial centroid selection, which can affect the stability of results [6]. Additionally, the determination of the optimal number of clusters requires methodological rigor; in this regard, the improved method proposed by Wang Bingcan *et al.* [7] offers valuable guidance. Furthermore, influenza-related data are typically high-dimensional and exhibit strong spatiotemporal correlations, which raises additional challenges regarding the selection of appropriate features and distance metrics. Practical enhancements to the K-means algorithm in other fields, as exemplified by the work of He Meng [8] and Liu Chunyu [9], also offer valuable references for optimizing clustering performance in public health applications.

In conclusion, the K-means clustering algorithm provides an effective tool for the spatial analysis of influenza incidence. Previous studies have affirmed its value in identifying high-risk regions and formulating optimized control strategies. Nevertheless, future research should prioritize algorithmic refinement and the integration of multidisciplinary data to further advance its applicability in the domain of public health.

2. Methodology and Modeling

2.1. Principles of the K-Means Algorithm

The K-means algorithm, a classical unsupervised machine learning method also known as K-average or K-means clustering, is widely used for partitioning datasets into meaningful groups based on similarity. Its core principle is to iteratively partition n data samples into k clusters, ensuring high intra-cluster similarity and significant inter-cluster differentiation (Liu Fugang [5]). This is achieved by minimizing the sum of squared distances between each sample and the centroid of its assigned cluster, which ensures clusters are as compact and distinct as possible.

The algorithm proceeds in four key steps: Randomly select k data points as initial centroids, each representing the center of a cluster; assign each remaining data point to the cluster with the nearest centroid, typically using Euclidean distance as the similarity metric; recalculate the centroid of each cluster as the mean of all data points within that cluster; repeat the allocation and update steps until the objective function stabilizes (*i.e.*, the difference in sum of squared errors between iterations is below a threshold or the maximum number of iterations is reached).

As an unsupervised learning technique, K-means identifies latent natural groupings in data without relying on predefined labels, making it suitable for exploratory research. However, it has limitations: it requires predefining the number of clusters k, is sensitive to initial centroid selection (which may affect result stability [6]), and performs poorly with non-convex clusters, imbalanced cluster sizes, or noisy data.

This algorithm's emphasis on hierarchical structuring ensures rational classification while preserving the internal consistency of each group, yielding stable clustering outcomes (Liu Fugang [5]). Its simplicity, speed, and scalability for large datasets make it a practical choice for analyzing patterns in complex data, such as the urban characteristics and influenza incidence data in this study.

2.2. Determination of the Number of Clusters

The elbow method is employed to determine the optimal number of clusters *k*. This method involves computing the sum of squared errors (SSE or *E*) for different values of *k*, and plotting the *k*-*E* curve. The point at which the curve forms a distinct "elbow" is considered the optimal value of *k*. The specific steps are as follows: Run the K-means algorithm for $k = 1, 2, 3, \ldots, m$ (m is the preset maximum cluster number) respectively;

Calculate the square error and E corresponding to each *k*;

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \left\| x - \mu_i \right\|^2$$

Select the slope in the curve from the large to the small turning point as the optimal *k*.

In this experiment, combined with the characteristics of the data and the actual needs, the value range of the preset *k* is 2 to 5, and the city is finally divided into 3 categories through the elbow law (**Figure 1**).

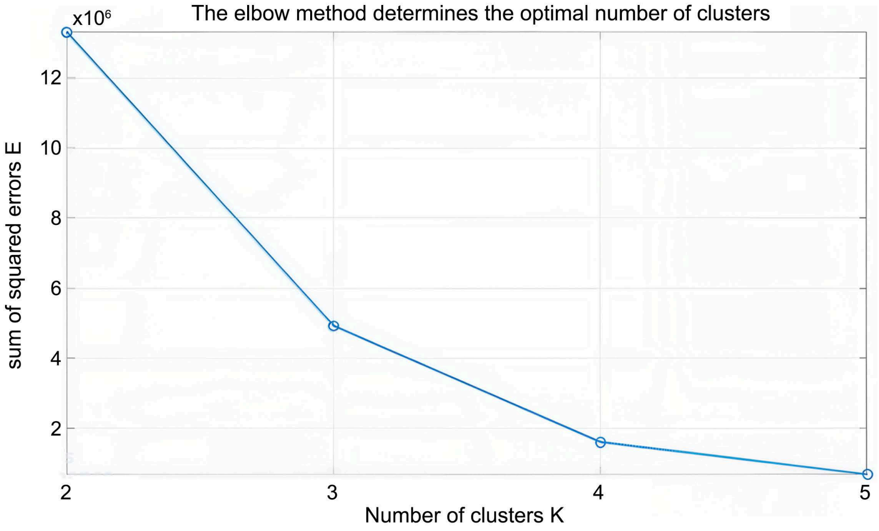

Figure 1. *k*-*E* curve.

2.3. Experimental Methods and Processes

1) Initialization: randomly select 3 city samples as the initial cluster center;

2) Allocation of samples: calculate the distance from each city sample to the center of each cluster, and distribute it to the nearest cluster;

3) Update the cluster center: for each cluster, calculate the average value of all urban samples assigned to the cluster as the new cluster center;

4) Convergence judgment: Calculate the new square error and *E*. If the difference between E and the previous iteration is less than the preset threshold (or the number of iterations reaches the upper limit), the algorithm will be terminated; otherwise, return to step 2 to continue the iteration.

3. Experimental Process

3.1. Data Preparation

The experimental data includes the urban line level, GDP and influenza incidence of 29 cities [1 Data source: China Statistical Yearbook and China Health Statistics Yearbook]. The data is as follows (**Table 1**):

Table 1. Data tables of 29 cities.

Urban-level	GDP (billion yuan)	Influenza incidence (/100,000)
1	4283	0.05
1	7450	0.11
1	4116	6.25
1	3423	6.25
2	3450	0.49

Continued

2	2932	0.14
2	2665	10.71
2	2515	8.04
2	1910	4.00
2	1956	0.26
2	2186	14.41
2	1901	0.02
3	1535	0.02
3	1680	0.05
3	1096	2.87
3	1619	0.17
3	2164	0.17
3	1134	1.51
3	1548	1.91
3	883	1.91
3	770	3.49
3	1555	0.17
3	942	0.01
3	589	17.72
3	253	1.54
4	189	0.51
4	444	14.60
4	175	4.26
4	484	0.51

3.2. Data Processing

Use MATLAB software to process data. This experiment divides cities into three categories, uses the K-means function of MATLAB for cluster analysis, and then draws three-dimensional scatter plots to visually display the classification results.

4. Experimental Results and Analysis

4.1. Present Cluster Results through Visualization Technology

Through the three-dimensional scatter plot drawn by MATLAB, cities are divided into three categories, which are represented by blue, black and yellow, respectively. Each type of city shows different characteristics in terms of urban line level, GDP and influenza incidence (Figure 2).

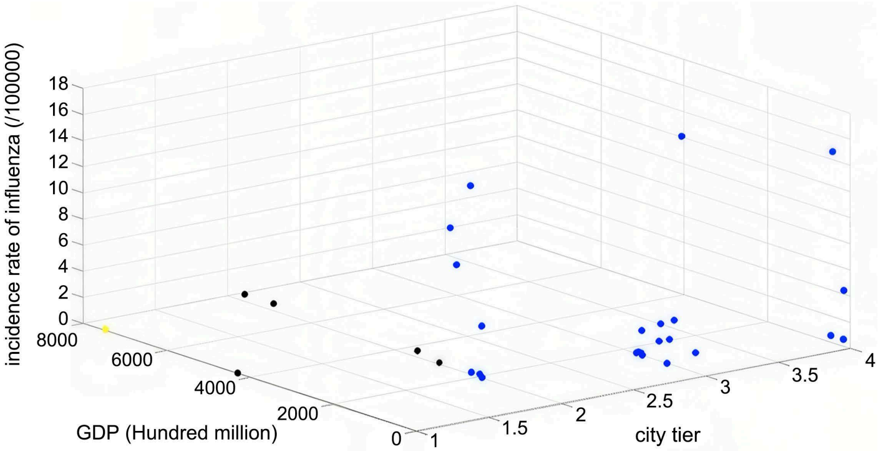

Figure 2. Data clustering result chart.

Through the three-dimensional scatter plot drawn by MATLAB, cities are divided into three categories, which are represented by blue, black, and yellow, respectively.

Classification results:

The cities represented by the yellow dots are "low incidence, high GDP and low-level" cities. GDP is nearly 800 billion, the urban level is close to 1, and the incidence of influenza is close to 0.

Cities represented by black dots: urban groups with the core characteristics of "wide GDP range, low level and low incidence", including areas with different economic scales, low urban level and stable public health performance. GDP is mostly in the range of 200 to 800 billion, concentrated in the urban line level from 0 to 2, and the incidence of influenza is 0/10 to 4/100,000.

The cities represented by the blue dot: covering urban groups with "high-level, wide GDP range and diverse incidence", including areas with different economic scales but 2.5+ at the urban level, with large differences in public health performance. The urban line level is mostly 2.5 to 4, the GDP span is large, and the incidence of influenza is 0/100,000 to 16/100,000.

Inter-class and intra-class similarities and differences:

1) Between categories (different colors)

Similarities: They are all urban samples, based on the three-dimensional characteristic classification of "GDP, urban line level and influenza incidence", reflecting the comprehensive performance of the city in economic, hierarchical and public health.

Differences:

Distribution range: the yellow class is isolated, the black class is concentrated in the "low-level and broad GDP range", and the blue class is concentrated in the "high-line and medium-low GDP".

2) In the class (similar colors)

Similarities:

Black class: "Low-line level, wide GDP and low incidence" characteristics overlap, spatial distribution concentration.

Blue class: "high line level" features are unified and gather around the line level.

Differences:

Black category: GDP fluctuates significantly, but the overall range is relatively narrow; the incidence rate is low, and there is no significant difference.

Blue category: GDP span is very large, incidence varies greatly, and internal characteristics are discrete.

4.2. Carry out an In-Depth Analysis of the Causes of the Phenomenon

Urban level and GDP: Cities with lower urban levels have relatively low GDP; cities with higher urban levels have higher GDP.

Urban and influenza incidence: cities with lower urban levels have a higher incidence of influenza; cities with higher urban levels have a lower incidence of influenza.

GDP and influenza incidence: Cities with relatively low GDP have a higher incidence of influenza; cities with high GDP have a lower incidence of influenza.

There is a correlation between the urban level, GDP and the incidence of influenza. When the urban level is low, the GDP is often relatively low, and the incidence of influenza is high; while when the urban level is high, the GDP is usually high, and the incidence of influenza is low.

For cities in the "high-hierarchy, high-GDP, low-incidence" cluster (blue cluster), these cities have a relatively sound medical resource base. The key is to maintain their existing advantages in prevention and control and, at the same time, leverage their regional radiating role. Some redundant resources (such as advanced detection equipment and professional prevention and control teams) can be allocated to surrounding low-hierarchy cities. Strengthen the construction of inter-regional joint prevention and control mechanisms, and drive the improvement of prevention and control capabilities in surrounding areas through technology transfer and personnel training.

For cities in the "low-hierarchy, wide-range GDP, low-incidence" cluster (black cluster), although the incidence rate in these cities is low, potential risks brought by economic differences need to be watched out for, and resource shortages that may lead to a rebound of the epidemic should be avoided. For cities with lower GDP, increase investment in basic medical facilities, such as building standardized community health service centers and stockpiling basic epidemic prevention materials. For cities with higher GDP, focus on improving the early warning system and use their economic advantages to establish a rapid response mechanism.

For cities in the "low-hierarchy, low-GDP, high-incidence" cluster (high-incidence groups outside the yellow cluster), these cities are weak links in prevention and control. Priority should be given to filling resource gaps to reduce the risk of epidemic spread. The central or provincial finance should increase the intensity of transfer payments and invest medical resources in a targeted manner, such as increasing the number of infectious disease beds and equipping vaccine cold storage equipment. Give priority to launching free influenza vaccination programs, covering key groups such as the elderly and children. Establish a counterpart support relationship with high-hierarchy cities and regularly dispatch medical teams to guide prevention and control work.

5. Conclusions

5.1. Discussion

In view of the incidence of influenza in different cities, this article collected data on the urban level, GDP and influenza incidence of 29 cities, and used the elbow method to determine the best cluster number as 3. On this basis, the sample data of 29 cities was clustered and analyzed using the K-means method. The results showed the correlation between the urban line level, GDP and influenza incidence of 29 cities: the incidence of influenza in high-line and high-GDP cities is generally low, while the incidence of low-line and low-GDP cities The incidence rate is relatively high, and similar cities have similar characteristics in terms of economic level, administrative level and public health performance, while there are obvious differences between different types of cities.

The essence of the urban line-level reflects the functional hierarchy of the city. High-level cities are usually regional medical centers with a high degree of concentration of resources, such as the concentration of three-A hospitals and a perfect disease control system, while low-level urban medical resources are scattered and have a weak foundation. Combined with the "negative correlation between line level and incidence" in the classification, one of the core intermediary factors that can be derived from this relationship is the distribution of medical resources—the higher the line level, the more sufficient the resources, and the stronger the ability to prevent and control influenza.

5.2. Causal Mechanisms

1) Medical resource distribution

High-level cities and those with robust GDP typically function as medical centers, characterized by concentrated resources such as tertiary hospitals, well-funded disease control agencies, and specialized healthcare personnel. These resources enable more effective surveillance, rapid response to outbreaks, and widespread access to preventive measures. In contrast, low-level, low-GDP cities often suffer from fragmented healthcare systems, inadequate funding for epidemic preparedness, and limited access to vaccines or antiviral treatments—factors that exacerbate transmission risk. For instance, the yellow cluster (low incidence, high GDP, low level) may represent exceptions where local economic strength compensates for administrative ranking, enabling investments in private healthcare or targeted public health programs.

2) Urban infrastructure and socioeconomic conditions

High-GDP cities generally boast superior infrastructure, including improved sanitation, ventilation systems in public spaces, and efficient transportation networks that facilitate healthcare access. These features reduce virus transmission pathways and ensure timely medical intervention. Additionally, socioeconomic factors such as education levels and public health awareness—often correlated with GDP—may play a role: residents in wealthier cities may be more informed about preventive behaviors, further lowering incidence rates.

5.3. Limitations of the Study

1) Sample size and representativeness

The analysis is based on 29 cities, a relatively small sample that may not capture the full diversity of urban contexts. This limits the ability to generalize conclusions to broader regional or national scales. A sample of this scale, when considered against the vast and varied landscape of urban environments-encompassing metropolises, mid-sized cities, small towns, and even rural-urban hybrids that blur the line between urban and rural characteristics-struggles to encapsulate the full spectrum of contextual diversity. For instance, cities with unique climatic profiles, such as those in high-altitude regions with extreme temperature variations or coastal areas with high humidity, may exhibit distinct influenza transmission dynamics shaped by local weather patterns. A sample of 29 cities cannot adequately account for these nuances, potentially leading to an oversimplification of how influenza incidence interacts with local demographics.

2) Algorithm limitations

The application of the algorithm also has certain limitations. For example, the cluster number k must be given in advance before clustering, and the algorithm is sensitive to the initial value. Different initial values may lead to different results. It is not suitable for clusters with non-convex shapes or clusters with large differences in size. In particular, the K-means algorithm is sensitive to “noise” and isolated point data. This prompts us to see that for particularly complex problems or data, we need to choose other clustering methods and algorithms for analysis.

5.4. Directions for Future Research

1) Expand sample size and scope

Include a larger, more diverse set of cities and incorporate rural areas to compare urban-rural dynamics. Longitudinal data would also help track how incidence patterns evolve with changes in urban development or policy interventions.

2) Refine analytical methods

Use advanced clustering algorithms such as DBSCAN and hierarchical clustering to handle non-convex or unevenly sized clusters, and apply causal inference techniques like instrumental variables to disentangle correlation from causation.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Han, X.C., Hu, Y.W., Wu, Q.Y., Hu, M. and Zeng, S.Y. (2024) Personnel Management Abnormal Data Identification and Automatic Processing System Based on K-Means Clustering Algorithm. *Electronic Design Engineering*, **32**, 27-31.

[2] Weng, Z.Y. (2025) Automatic Detection System of DC Power Grid Transformer Fault Based on K-Means Clustering Algorithm. *Automation and Instrumentation*, **40**, 86-90.

[3] Bao, W.N. (2018) Study on the Genomic Variation Characteristics of Influenza A H1N1 Virus. Master's Thesis, Shijiazhuang Railway University.

[4] Xia, H. and Zhu, H. (2024) Explore the Application of Artificial Intelligence in the Early Warning System of Infectious Diseases in Hospitals. *China Journal of Health Information Management*, **21**, 571-577.

[5] Liu, F.G. (2023) Applied Research on K-Means Clustering Algorithm in Network Security Detection. *Journal of Suihua College*, **43**, 157-160.

[6] Zhou, X.D., Dong, H.Q., Zhang, K.P., Hou, J.C. and Sun, S.F. (2025) Research on K-Means Initial Cluster Center Optimization Algorithm Based on Geometry. *Instrument Technology*, No. 2, 66-69, 73.

[7] Wang, B.S., Wang, G.C. and Wei, Y.H. (2025) Improvement of the Cluster Number Determination Method Based on K-Means Algorithm. *Statistics and Decision-Making*, **41**, 59-64.

[8] He, M. (2024) Network Anomaly Data Mining and Classification Method Based on Improved K-Means Clustering Algorithm. *Wireless Interconnection Technology*, **21**, 119-122.

[9] Liu, C.Y. (2025) Distributed Energy Storage Cluster Division Method Based on Improved K-Means Clustering Algorithm. *Northeast Power Technology*, **46**, 1-5.

Appendix

Experimental code

```
% Define the data matrix
a = [1 4283 0.05;
    1 7450 0.11;
    1 4116 6.25;
    1 3423 6.25;
    2 3450 0.49;
    2 2932 0.14;
    2 2665 10.71;
    2 2515 8.04;
    2 1910 4.00;
    2 1956 0.26;
    2 2186 14.41;
    2 1901 0.02;
    3 1535 0.02;
    3 1680 0.05;
    3 1096 2.87;
    3 1619 0.17;
    2 2164 0.17;
    3 1134 1.51;
    3 1548 1.91;
    3 883 1.91;
    3 770 3.49;
    3 1555 0.17;
    3 942 0.01;
    3 589 17.72;
    3 253 1.54;
    4 189 0.51;
    4 444 14.60;
    4 175 4.26;
    4 484 0.51];
% Use K-Means algorithm for clustering
b=kmeans(a,3);
% Draw a three-dimensional scatter plot
[m n]=size(a);
for ii=1:m
  a(ii,4)=b(ii);
  switch b(ii)
    case 1
      scatter3(a(ii,1),a(ii,2),a(ii,3),'b','filled')
      hold on
    case 2
```

```
            scatter3(a(ii,1),a(ii,2),a(ii,3),'k','filled')
            hold on
        case 3
            scatter3(a(ii,1),a(ii,2),a(ii,3),'y','filled')
            hold on
    end
end
xlabel('City Line Level');
ylabel('GDP (billion yuan)');
zlabel('Influenza incidence (/100,000)');
```