International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 2, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v2n1.48
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Visual Analysis of Epidemic Epidemiological Investigation Track Data

Dan Luo[1], Yadong Wu[2,*], Kai Liu[2], Weihan Zhang[2]

[1]School of Automation and Information Engineering, Sichuan University of Science & Engineering, Yibin, China

[2]College of Computer Science and Engineering, Sichuan University of Science & Engineering, Yibin, China

ABSTRACT

The visualization analysis of the flow data of the novel coronavirus pneumonia epidemic can intuitively show the development dynamics of the epidemic, explore the law of epidemic transmission, and provide a new way of thinking and method for the analysis of the epidemic situation. Firstly, the convection-modulated trajectory data is preprocessed, and then the convection-modulated data is theoretically transferred on the basis of expectation maximization statistics to realize multi-trajectory fusion algorithm, and finally the trajectory fitting of infected population is carried out. Finally, the multi-type thematic map is used to combine the flow data with the map system to realize the visualization of flow data and conduct in-depth analysis, so as to dig out the distribution characteristics of the flow data, provide a reference for the prevention and control of the novel coronavirus epidemic in a place, enable people to correctly understand the spread of the epidemic, and promote the development of preventive medicine.

KEYWORDS

Expectation-Maximization; Epidemiological Survey Data; Visualized Analysis; Trajectory Visualization

1. INTRODUCTION

At present, the incidence of all kinds of epidemics is increasing[1] year by year, of which the novel [2]coronavirus disease is the most serious public health event since the 21st century, with the emergence of various "[3]variant of concern (VOC)". The Delta strain[4], for example, has not only caused serious harm to human health, but also caused huge losses [5]to the global economy. Therefore, it is essential to analyze and summarize the epidemic, among which the data analysis of epidemiological survey (hereinafter referred to as "flow survey") is extremely important.

Various research centres, federal and provincial websites provide multiple forms of visualization. Most of these visualizations focus on one part of the message. For example, the health big data platform of the regional health information system in Ningbo, Zhejiang province, has seen initial results in epidemic prevention and control[6]. Johns Hopkins University (JHU) focuses on global and US trends[7], while the dashboard [8]created by the World Health Organization (WHO) focuses on global trends, but there is a greater need for a visualization that can provide more information than trends, such as the analysis of convection-based data to help track the spread[9] of the virus.

Traditional stream-modulated analysis methods generally obtain relevant data through manual survey methods such as home visit surveys and telephone inquiries. However, under the computer-aided calculation, the diagnostic data, OD matrix and other results are obtained by statistical analysis

method and corresponding mathematical model. Improving the mathematical model and statistical methods can improve the quality of analysis results, but these investigation and analysis methods still have the following deficiencies:

(1) The flow modulation data is large in quantity and low in value density, and it is difficult to dig its inherent law in the face of massive data. The trajectory data ignores the attribute information, making the clustering logic not rigorous enough.

(2) In the system, due to the lack of support for the corresponding spatiotemporal model, the time attribute is rarely considered in the analysis process, and the unified spatiotemporal system is not studied to diagnose the trajectory activities of individuals.

In view of the above problems, this paper carried out a study on visual clustering and trajectory fitting of epidemic flow data. Firstly, relevant data of epidemic flow monitoring in Zhengzhou were collected, and the obtained flow monitoring data were based on multi-track fitting of EM algorithm to demonstrate the effects of visual trajectory data and trajectory animation, and provide more detailed flow monitoring information display.

Epidemic data covers medical, communications, information media, information technology, and other fields, and the data formats are different for different fields. Therefore, according to the fields involved in epidemiological data, epidemiological data is divided into four dimensions: spatio-temporal dimension, semantic level, transmission model, and relevant contextual content.

The spatio-temporal dimension includes temporal attributes and spatial attributes. Spatiotemporal data is a common data used in current visual analysis methods of epidemic transmission. Semantic information is mainly based on text data, such as short text data sets[10] related to epidemic data in news media or journals, which can visually analyze the content of epidemic text. Epidemic model data includes simulated simulation data, model-related parameters, etc. Epidemic model includes transmission dynamics model and individual model; The general epidemic model is divided into SI model, SIR Model, SIRS model and SEIR model according to the epidemic type. [11]According to the transmission mechanism, it can be divided into different types based on ordinary differential equation, partial differential equation and network dynamics. These epidemic model data can simulate and predict the development trend of different types of epidemics. The epidemiological context contains data such as virus genes and popular science information[12]. The presentation of epidemic-based images, such as posters of viruses, dynamic information content displays and genetic maps, accelerates the public's understanding of the epidemic context.

According to epidemic data, the tasks of epidemic visual analysis can be divided into the following four categories: spatial and temporal visual analysis of epidemic data, focusing on the distribution and spread[13] of epidemics; Semantic visual analysis of epidemic data, which mainly mines the content of epidemic text data and analyzes emotional changes; Visual analysis of epidemic models based on epidemic model data, and visual analysis of information maps that use data modeling to analyze the direction of transmission or predict the progress of an epidemic, require less visual visual representation of the analysis content based on the content of the interpreted data. Depending on the analysis task, different visualization methods can be applied to better present the epidemic data.

This paper mainly studies how to select appropriate trajectory fitting processing methods to process the flow modulation data, and uses visualization technology to display the results of flow modulation data processing. Therefore, this paper involves two research directions: trajectory fitting research and visualization of fitting results research.

2. RESEARCH STATUS OF VISUAL ANALYSIS OF TRAJECTORY DATA

There are usually three methods for visual analysis of trajectory data: direct visualization, aggregate visualization, and feature visualization. Direct visualization is the most basic visual analysis method,

displaying track data in order for the user to observe. Aggregate visualization first calculates the aggregate data of the trajectory and then plots it out. Feature visualization first calculates the features of the trajectory and then plots them using direct or aggregate methods.

2.1. Direct Visualization

In this approach, the computer is the main star of the "visual" part, while the "analysis" is mostly performed by humans. Direct visualization methods can be further divided into position animation[25], path visualization[26], space-time cube[27], timeline visualization[28], and parallel coordinates[29].

2.2. Aggregation Visualization

When the trajectory data is large, aggregate visualization is generally considered. The aggregation calculation of trajectory data is closely related to the spatial data cube[31] in data mining, both of which are based on multidimensional data models to carry out statistical analysis of data of each dimension. Based on the different[30] selected dimensions, clustering visualization can be divided into three categories: spatio-temporal and attribute clustering[32][25], start-destination clustering[33] and path clustering.

2.3. Feature Visualization

When the features of concern are relatively determined and can be calculated, the feature visualization method is used. This method is subdivided into event visualization and pattern visualization. The time span that meets certain conditions is called event[34]. Event visualization focuses on partial trajectories and related events that meet certain conditions. Trajectory events are all spatial events. Pattern visualization can focus on both general and behavioral features[35]. General features are applicable to most trajectory data, but behavioral features are only applicable to specific applications.

2.4. Summary

Advantages and disadvantages of the three visualizations：

The advantages of the direct visualization are good tolerance of noise and outliers in data; Suitable for exploratory analysis; Accurately retain the information in the data; The method is simple. The cons of the direct visualization are not suitable for analysis of a large number of tracks; It is not systematic enough and will miss many features; The task of manual analysis is heavy; The analysis process is difficult to reproduce and the results are difficult to evaluate.

The advantages of the aggregation visualization are visual analysis that supports a large number of tracks; Can directly answer questions about aggregation characteristics; The pressure of manual analysis is reduced. The cons of the aggregation visualization are aggregated data is less intelligible;

Easy to discard information; It is difficult to study the interaction and relative motion[35] between trajectories; Additional programming implementations are required.

The advantages of the feature visualization are support visual analysis of a large number of tracks;

Features can be studied directly; It can be searched automatically by computer; Manual analysis is less stressful. The cons of the feature visualization are a lot of information unrelated to features is lost; Poor support for more exploratory tasks; Feature calculation requires a lot of programming.

In this paper, trajectory fitting is carried out based on massive flow modulation trajectory data, that is, theoretical migration is carried out on the basis of expectation maximization (EM algorithm) statistics, multi-trajectory fusion algorithm is realized, and finally trajectory fitting of infected population is carried out to achieve visual analysis effect. Provide help for users to understand the

epidemic information and enhance their awareness of epidemic prevention. In general, trajectory data has characteristics of spatio-temporal sequence, cross-frequency sampling and poor data quality. By analyzing the trajectory data of the flow modulation, the moving characteristics of the active areas and active scenes of the confirmed patients can be mined. The mining and analysis of the flow modulation trajectory data is the key research object of the researchers.

3. DATA SOURCES

Data source for this article is: Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The main experimental data was the trajectory data of a confirmed patient in a certain place. Specific data items included patient number, date, location name, longitude and latitude coordinates (POI address resolution) and detailed trajectory description,, as shown in Fig. 1 and Fig. 2.

id	data	place	lon	lat
1	2020-01-07之前	Place1	114.311582	30.598467
1	2020-01-07	Place2	114.709823	34.130203
1	2020-01-08	Place3	114.879742	34.040425
1	2020-01-10	Place4	113.753865	34.78167
1	2020-01-20	place5	113.664667	34.718381

Figure 1. Infographic of patient trajectory data

Case 42: Female, 50 years old, current address 1. On the evening of January 22, I took the high-speed train to Address 2, on January 23, I took a bus from Address 2 to Address 3, on January 24, I took a local taxi to address 4, and I was diagnosed on January 29.

Case 43: Female, 26 years old, current address 1. On January 22, he took a train from address 2 to address 1, then took bus 81 home, went to address 3 in the afternoon of January 26, and was diagnosed on January 29.

Case 44: Female, 47 years old, current address 1. On January 19, he drove to address 2 to visit his relatives and came into contact with suspected patients. On the morning of January 26, he drove from address 2 to address 1. On January 28, he was sent by 120 emergency vehicle to address 3 for treatment.

Figure 2. Detailed track description diagram

In addition to patient information, hospitals are important institutions related to this epidemic. Therefore, this project also uses python language and Baidu map API interface to crawl POI points of hospitals and delete irrelevant institutions. The hospital-related information obtained is shown in Fig. 3.

1	jinkaiqu	113.8019	35.18549
2	jinkaiqu	113.8019	35.18549
3	jinkaiqu	113.8019	35.18549
4	jiulongzhen	113.8582	34.68798
5	zhengzhoudaxue	113.627	34.82698
6	jinkaiqu	113.8019	35.18549
7	zhengzhoushi	113.7472	34.71997
8	luoheshi	114.0312	33.57939
9	henansheng	112.1105	32.67698
10	jinshuiqu	113.6547	34.79294

Figure 3. Is the information graph of POI points in hospitals

Because the data is natural language, it is difficult for computer to recognize, so it is necessary to adjust the data structure and convert it into semi-structured data or structured data, so as to facilitate the identification and display of trajectory points. Therefore, all the patient track information is extracted, including patient ID, date, location name three data items, stored in excel file.

Based on the above data analysis, in view of the absence, duplication, anomaly and other situations in the process of data collection and collation, so to achieve the standardized storage, management and application of the data, we need to adopt the corresponding data processing method according to the characteristics of the original data, and store and standardize the operation in accordance with the specified data table structure and data type. The specific implementation is as follows:

(1) Vacancy value processing

Determine whether there is a missing value, if there is a corresponding method to supplement according to the data characteristics.

(2) Abnormal data clearing

Set the specified range of attribute values, and make average correction to abnormal data.

(3) The removal of duplicate data

Delete the duplicate data in the data set and keep only one copy of it, thus eliminating redundant data.

(4) Format standardization

The unit, type and size of the data are standardized, and the meaningless characters are deleted.

Due to the large amount of data of track points, it is time-consuming and labor-intensive to fill in longitude and latitude manually, so Geocoder() of Baidu map API is used for address resolution to obtain the longitude and latitude coordinates of all points, as shown in Fig. 4.

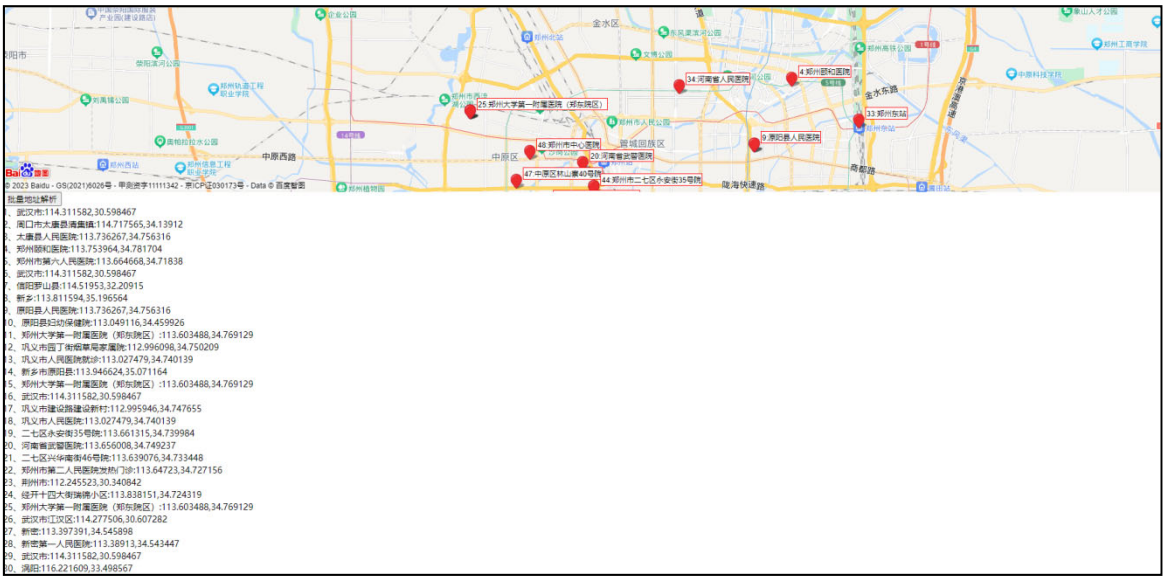

Figure 4. Address resolution result diagram

(5) Data cleaning

Because the locus points of some patients are distributed abroad, some are distributed in other provinces or surrounding cities, and some addresses have some "same name" phenomenon, when using Baidu API for address resolution, the definition of the range is set to a place, there will be a part of the locus point location error, so it is necessary to modify the longitude and latitude of the analysis error point by using the "Baidu coordinate system".

In order to better discover the way of transmission, the activity trajectories of infected people should be fully excavated. Due to the discrete nature of the continuous movement path, the uncertainty of sampling accuracy, location and the influence of pre-processing methods, the analysis of trajectory data has brought certain difficulties. Based on the expectation maximization (EM algorithm) statistics, the theoretical migration is carried out to realize the multi-trajectory fusion algorithm, and the trajectory fitting of the infected population is carried out. Show the effect of visual trajectory data and trajectory animation, and visually present the flow modulation data.

4. RESEARCH METHODS

This study focuses on the visual analysis of epidemic data in a certain place and the study of the transmission chain of the epidemic, which is of great significance for the effective prevention and suppression of the epidemic and the protection of people's life safety. Use a variety of intuitive and interactive visualizations to visualize the outbreak through patient trajectory data.

Trajectory fitting with convective data. In the actual data processing, it is necessary to make the error of the measured value tend to zero with repeated measurement, and use the arithmetic average idea to correct the error of the obtained value for many times to obtain the accurate GPS measured data points of the track line, and then fit the processed GPS measured data points of the precise track line to generate the track fitting curve. Based on the expectation maximization (EM algorithm) statistics, the theoretical transfer is carried out to realize the multi-track fusion algorithm, and finally the trajectory fitting of the infected population is carried out.

Using visual analysis, a reasonable interactive exploration method is designed to help users understand the epidemic information, learn relevant epidemic prevention and control knowledge, and enhance users' awareness of epidemic prevention.

This project developed a visualization and analysis system for epidemic flow survey data on the map system platform. The visual analysis and modeling of epidemic flow analysis data is accomplished

through the combination of various components. Almost all epidemic data is related to geographical distribution. The data is first structured and then stored in a database. Each data should have corresponding map location parameters, which mainly include epidemic information, time and map location attributes. The overall architecture of the system is shown in Figure 5. Firstly, the text trajectory information data is obtained, and the original data is pre-processed; Secondly, cluster analysis is carried out on the convective trajectory data; Finally, the trajectory fitting results are obtained, and the results are displayed by visual analysis.

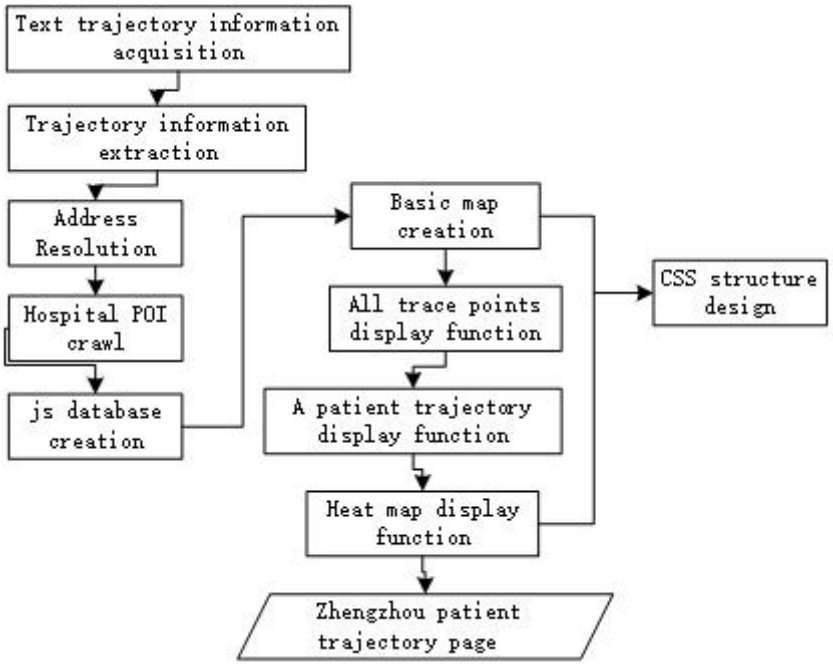

Figure 5. Research framework for visual analysis of epidemic flow control trajectory data

Due to the influence of noise and other factors, the flow-modulated data used in this paper has certain errors, and the trajectory fitting of the flow-modulated data is easy to cause problems such as track segmentation and track information loss. Under this background, this paper puts forward a macro trajectory fitting idea, focusing on the trajectory fitting problem. Traditional trajectory fitting can be divided into three types: modern intelligent algorithms, parameter estimation algorithms and statistical methods. Therefore, this paper proposes a theoretical migration based on the expectation maximization (EM algorithm) statistics to realize a multi-trajectory fusion algorithm.

4.1. EquationsEM Algorithm

EM algorithm solves the following problems: given input observation variable data Y, hidden variable data Z joint distribution, conditional distribution; $\mathrm{P}(Y,Z|\theta)\,\mathrm{P}(Y,Z|\theta)$ And estimate the output model parameter θ. For the input model, using the maximum likelihood criterion, the objective function can be established:

$$L(\theta)=\lg \mathrm{P}(Y|\theta)=\lg \sum_{Z} \mathrm{P}(Y,Z|\theta)=\lg\left[\sum_{Z} \mathrm{P}(Y,Z)\mathrm{P}(Z|\theta)\right] \tag{1}$$

For the above optimization problem, under the condition of complete observed variables, the criterion function can be solved by the maximum likelihood estimation (MLE) criterion. $L(\theta)$ But in the presence of hidden variable Z, the criterion function has no closed solution. It can be regarded as known, and the parameter θ is further solved, and the solution is completed after repeated iteration. $\mathrm{P}(Z|\theta)$However, this method has the following two shortcomings: in the process of MLE solving, there are summation of numerators and denominators and integral terms, and repeated iterations will

cause the operation to become complicated; Can not meet the convergence conditions. Solid needs to use the EM algorithm, considering that the criterion function after the i th iteration is not less than the original criterion function, according to JENSEN inequality:

$$\begin{aligned} L(\theta)-L(\theta^{(i)}) &= \lg\left[\sum_Z \mathrm{P}(Y|Z,\theta^{(i)})\frac{\mathrm{P}(Y|Z,\theta)\mathrm{P}(Z|\theta)}{\mathrm{P}(Y|Z,\theta^{(i)})}\right]-\lg \mathrm{P}(Y|\theta^{(i)}) \\ &\geq \sum_Z \mathrm{P}(Y|Z,\theta^{(i)})\lg\left[\frac{\mathrm{P}(Y|Z,\theta)\mathrm{P}(Z|\theta)}{\mathrm{P}(Y|Z,\theta^{(i)})}\right]-\lg \mathrm{P}(Y|\theta^{(i)}) \\ &= \sum_Z \mathrm{P}(Y|Z,\theta^{(i)})\lg\left[\frac{\mathrm{P}(Y|Z,\theta)\mathrm{P}(Z|\theta)}{\mathrm{P}(Y|Z,\theta^{(i)})\mathrm{P}(Y|\theta^{(i)})}\right] \end{aligned} \tag{2}$$

$\theta^{(i)}$ The maximum value of the constant term can be omitted, further obtained from the above equation:

$$\begin{aligned} \theta^{(\mathrm{i}+1)} &= \arg\max_\theta \sum_Z \mathrm{P}(Y|Z,\theta^{(\mathrm{i})})\lg(\mathrm{P}(Y|Z,\theta)\mathrm{P}(Z|\theta)) \\ &= \arg\max_\theta \sum_Z \mathrm{P}(Y|Z,\theta^{(\mathrm{i})})\lg(\mathrm{P}(Y,Z|\theta)) \\ &= \arg\max_\theta Q(\theta,\theta^{(i)}) \end{aligned} \tag{3}$$

EM algorithm specific steps are as follows:

Step 1: Select the initial parameter values and start iteration; $\theta^{(0)}$

Step 2: Step E (find $Q(\theta,\theta^{(i)})$): denoted $\theta^{(i)}$ as the i th iteration parameter θ estimate, in the i+1 iteration E step, calculate:

$$Q(\theta,\theta^{(\mathrm{i})}) = \mathrm{E}_Z\left[\lg \mathrm{P}(Y,Z|\theta)\middle|Y,\theta^{(\mathrm{i})}\right] = \sum_Z \lg \mathrm{P}(Y,Z|\theta)\middle|\mathrm{P}(Y,Z|\theta^{(\mathrm{i})}) \tag{4}$$

Where: is the conditional probability distribution of the implicit variable $\left|\mathrm{P}(Y,Z|\theta^{(\mathrm{i})})\right|\mathrm{P}(Y,Z|\theta^{(\mathrm{i})})\mathrm{Z}$ under the given observation data $\theta^{(i)}$ Y and the current parameter estimation.

Step 3: M step to maximize θ, $Q(\theta,\theta^{(\mathrm{i})})$ that is, on the premise that the conditional probability density of the hidden variable is given, MLE is used to achieve parameter estimation. Determine the parameter estimates for the i+1 iteration: $\theta^{(i+1)}$

$$\theta^{(\mathrm{i}+1)} = \arg\max_\theta Q(\theta,\theta^{(i)}) \tag{5}$$

Step 4: Repeat steps 2 and 3, meet convergence conditions end.

At this point, the whole derivation process of EM algorithm is completed.

4.2. Multi-trajectory fitting algorithm

(1) Mixed Gaussian model

Derived according to EM algorithm, solve for mixed Gaussian model (GMM). $\mathrm{P}(Z_\mathrm{j}\in Y_\mathrm{k}|Y_\mathrm{j},\ \Theta^{(\mathrm{i})})$ $Z_\mathrm{j}\in Y_k$ Denotes that the JTH observation point comes from the KTH model, and denotes the set of parameters. Θ K mixed Gaussian models can be used by the full probability formula, which gives:

$$\mathrm{P}\left(Z_{\mathrm{j}} \in Y_{\mathrm{k}} \middle| Y_{\mathrm{j}},\ \Theta^{(\mathrm{i})}\right)=\frac{\omega_{\mathrm{k}}^{(i)} f_k\left(Y_j \middle| Z_j \in Y_k, \theta_k\right)}{\sum_{k=1}^{K} \omega_{\mathrm{k}}^{(i)} f_k\left(Y_j \middle| Z_j \in Y_k, \theta_k\right)} \tag{6}$$

Further write the criterion function:

$$Q\left(\Theta, \Theta^{(\mathrm{i})}\right)=\sum_{j=1}^{N}\sum_{k=1}^{K} \lg\left(\omega_k\right) \mathrm{P}\left(Z_j \in Y_k \middle| Y_j, \Theta^{(i)}\right)+\sum_{j=1}^{N}\sum_{k=1}^{K} \lg\left(f_k\left(Y_j \middle| Z_j \in Y_k, \theta_k\right)\right) P\left(Z_j \in Y_k \middle| Y_j, \Theta^{(i)}\right) \tag{7}$$

Where: $\theta_k=\left[\mu_k, \sigma_k\right]$ is the parameter corresponding to distribution k; $\Theta=\left\{\theta_1,\ \theta_2, \ldots, \theta_{\mathrm{k}}\right\}$ is the parameter set; N is the number of samples.

First of all, the ω_{k} optimization, since, using the $\sum_{k=1}^{K} \omega_k=1$ Lagrange multiplier to solve:

$$J_\omega=\sum_{j=1}^{N}\sum_{k=1}^{K}\left[\lg\left(\omega_k\right) \mathrm{P}\left(Z_j \in Y_k, \Theta^{(i)}\right)\right]+\lambda\left[\sum_{k=1}^{K} \omega_k-1\right] \tag{8}$$

Partial differential solution:

$$\frac{\partial J_\omega}{\partial \omega_{\mathrm{k}}}=\sum_{J=1}^{N}\left[\frac{1}{\omega_{\mathrm{k}}} \mathrm{P}\left(Z_j \in Y_k \middle| Y_j, \Theta^{(i)}\right)\right]+\lambda=0 \tag{9}$$

Get:

$$\omega_{\mathrm{k}}^{(i+1)}=\frac{\sum_{J=1}^{N}\left[\mathrm{P}\left(Z_j \in Y_K \middle| Y_j, \Theta^{(i)}\right)\right]}{N} \tag{10}$$

Optimize the internal parameters to θ_k get the criterion function:

$$J_\Theta=\sum_{\mathrm{j}=1}^{N}\sum_{k=1}^{K} \lg\left(f_k\left(Y_j \middle| Z_j \in Y_k, \theta_k\right)\right) \mathrm{P}\left(Z_j \in Y_k \middle| Y_j, \Theta^{(i)}\right) \tag{11}$$

For Gaussian distributions:

$$f_k\left(Y_j \middle| Z_j \in Y_k, \theta_k\right)=\frac{1}{(2\pi)^{d/2}\left|\sum_k\right|^{1/2}} \bullet \exp\left[-\frac{1}{2}\left(Y_j-\mu_k\right)^T \sum_k^{-1}\left(Y_j-\mu_k\right)\right] \tag{12}$$

Derivation of the parameter:

$$\sum_k^{(i+1)}=\frac{\sum_{j=1}^{N} \mathrm{P}\left(Z_j \in Y_k \middle| Y_j, \Theta^{(i)}\right)\left(Y_j-\mu_k\right)^2}{\sum_{j=1}^{N} \mathrm{P}\left(Z_j \in Y_k \middle| Y_j, \Theta^{(i)}\right)} \tag{13}$$

$$\mu_k^{(i+1)}=\frac{\sum_{j=1}^{N} Y_j \mathrm{P}\left(Z_j \in Y_k \middle| Y_j, \Theta^{(i)}\right)}{\sum_{j=1}^{N} \mathrm{P}\left(Z_j \in Y_k \middle| Y_j, \Theta^{(i)}\right)} \tag{14}$$

At this point, the entire solution of GMM is completed.

(2) Mixed Laplacian model

Due to different mixing models, the parameter forms are also different. Because GMM is an even power solution is relatively simple, but the derivative of odd power has a symbolic function, can not directly derivative iteration, solid solution of odd power needs to further improve the universality of the mixed model (not limited to GMM, Laplace model). For the Laplacian distribution:

$$f(Y)=\frac{1}{2d}e^{-\frac{|Y-\mu|}{b}} \tag{15}$$

Where: μ is the mean value; b is the steepness coefficient.

For the mixed distribution of K models:

$$\mathrm{P}(Y_j|\Theta)=\sum_{k-1}^{K}\omega_k f(Y_j|\mu_k,b_k) \tag{16}$$

The E step is the same as the GMM solution, which is applicable to any mixed model. The same is true for solving the coefficients in step M, for solving the steep coefficients b:

$$\mathrm{b}_{\mathrm{k}}^{(i+1)}=\frac{\sum_{j=1}^{N}\mathrm{P}\left(Z_j\in Y_k\middle|Y_j,\Theta^{(i)}\right)\middle|Y_j-\mu_k\middle|}{\sum_{j=1}^{N}\mathrm{P}\left(Z_j\in Y_k\middle|Y_j,\Theta^{(i)}\right)} \tag{17}$$

To solve the mean:

$$\frac{\partial J_\Theta}{\partial\mu_k}=0\Rightarrow\sum_{j=1}^{N}sign(Y_j-\mu_k)\mathrm{P}\left(Z_j\in Y_k\middle|Y_j,\Theta^{(i)}\right)=0 \tag{18}$$

At this time, iterative solution cannot be completed, and in the final state of iteration, the parameters of order i can be considered to be approximately equal to the parameters of order i+1, thus the derivation result above is converted to:

$$\sum_{j=1}^{N}\frac{(Y_j-\mu_k)}{|Y_j-\mu_k|}\mathrm{P}\left(Z_j\in Y_k\middle|Y_j,\Theta^{(i)}\right)=0\Rightarrow\sum_{j=1}^{N}\frac{\left(Y_j-\mu_k^{(i+1)}\right)}{\left|Y_j-\mu_k^{(i)}\right|}\mathrm{P}\left(Z_j\in Y_k\middle|Y_j,\Theta^{(i)}\right)=0 \tag{19}$$

Thus completing the iteration of the mean:

$$\mu_k^{(i+1)}=\frac{\sum_{j=1}^{N}\frac{Y_j}{\left|Y_j-\mu_k^{(i)}\right|}\mathrm{P}\left(Z_j\in Y_k\middle|Y_j,\Theta^{(i)}\right)}{\sum_{j=1}^{N}\frac{1}{\left|Y_j-\mu_k^{(i)}\right|}\mathrm{P}\left(Z_j\in Y_k\middle|Y_j,\Theta^{(i)}\right)} \tag{20}$$

At this point, the parameter solution of the odd-power mixed model is completed.

(3) Mixed linear model

For different distributions, the noise characteristics of the radar signal received by the sensor are different. After solving the parameters based on most random stationary noise scenarios, we discuss how to migrate from the mixed distribution model based on EM to the trajectory fitting algorithm.

The maximum likelihood (MLE) algorithm can be used to fit a single trajectory. Multiple trajectories can not be directly applied for trajectory fitting. Suppose a bunch of data points, produced by 2 straight trajectories: (x_j,y_j)

$$\begin{cases} l_{1j} = a_1 x_j + b_1 + n_{1j} \\ l_{2j} = a_2 x_j + b_2 + n_{2j} \end{cases} \tag{21}$$

Where: n_{1j} and n_{2j} respectively are the corresponding random noise.

MLE can not be used directly to obtain parameters, and the noise of different trajectories can be regarded as the application of the mixed model, which is the mixed Gaussian model:

$$f_k\left(Y_j \middle| Z_j \in Y_k, \theta_k\right) = \frac{1}{(2\pi)^{d/2}\left|\sum_k\right|^{1/2}} \bullet \exp\left[-\frac{1}{2}\left(l_j - a_k x_j - b_k\right)^T \sum_k^{-1}\left(l_j - a_k x_j - b_k\right)\right] \tag{22}$$

Can be considered to be in $l_j - a_k x_j$ GMM, is. Y_j b_k μ_{k} Directly apply the GMM iteration result:

$$\sum_k^{(i+1)} = \frac{\sum_{j=1}^{N} \mathrm{P}\left(Z_j \in Y_k \middle| Y_j, \Theta^{(i)}\right)\left(Y_j - \mu_k\right)^2}{\sum_{j=1}^{N} \mathrm{P}\left(Z_j \in Y_k \middle| Y_j, \Theta^{(i)}\right)} \tag{23}$$

$$\mu_k^{(i+1)} = \frac{\sum_{j=1}^{N} Y_j \mathrm{P}\left(Z_j \in Y_k \middle| Y_j, \Theta^{(i)}\right)}{\sum_{j=1}^{N} \mathrm{P}\left(Z_j \in Y_k \middle| Y_j, \Theta^{(i)}\right)} \tag{24}$$

It can be seen that with more than one pair of solutions, a_k it is easy to get:

$$a_k^{(i+1)} = \frac{\sum_{j=1}^{N} x_j\left(l_j - \mu_k\right)\mathrm{P}\left(Z_j \in Y_k \middle| Y_j, \Theta^{(i)}\right)}{\sum_{j=1}^{N} x_j^2 \mathrm{P}\left(Z_j \in Y_k \middle| Y_j, \Theta^{(i)}\right)} \tag{25}$$

At this point, the theoretical derivation is complete.

The feasibility of linear multi-trajectory fitting is derived. The noise model can be derived from Gaussian model to other mixed models. For linear trajectories, it can also be deduced to various types of trajectories. More generally:

$$f_k\left(Y_j \middle| Z_j \in Y_k, \theta_k\right) = \frac{1}{(2\pi)^{d/2}\left|\sum_k\right|^{1/2}} \bullet \exp\left[-\frac{1}{2}\mathrm{g}\left(x_j \middle| \theta_{\mathrm{k}}\right)^T \sum_k^{-1} \mathrm{g}\left(x_j \middle| \theta_{\mathrm{k}}\right)\right] \tag{26}$$

g is a general expression, such as GMM is, and a more general $g = ax + b$ g can theoretically be any expression, as shown in Figure 6.

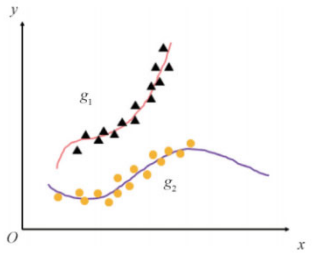

Figure 6. Schematic diagram of trajectory fitting

Just plug the concrete expression for g into the EM solution. The mixed model can theoretically achieve clustering of various shapes, and the noise can also be based on different distribution assumptions. The common k-means essence is the distribution assumption for the center point (cluster center); Gaussian mixture model is the distribution assumption for the line with slope 0 (the mean of GMM); The fitting of various trajectories is a general application of the EM algorithm.

5. VISUAL DESIGN

In this paper, two interactive analysis methods are designed to meet the needs of the analysis of the flow data, taking into account the characteristics of the data set: (1) Combined with the multi-scale analysis of the cumulative quantity distribution map of flow data, for the analysis of the overall situation; (2) the interactive exploration analysis method combined with map index is used for the feature analysis of flow analysis data.

5.1. Analysis of the overall situation

This study focuses on the visual analysis of epidemic data in a certain place and the study of the transmission chain of the epidemic, which is of great significance for the effective prevention and suppression of the epidemic and the protection of people's life safety. In this paper, the multi-scale cumulative quantity distribution map of streaming data (FIG. 10) is used to show the geographical distribution of the total number of epidemic data during the study period, and the cumulative quantity distribution of public opinion data is reflected from two scales by means of stratification and color. The study period begins on January 1, 2020 and ends on March 1, 2020.

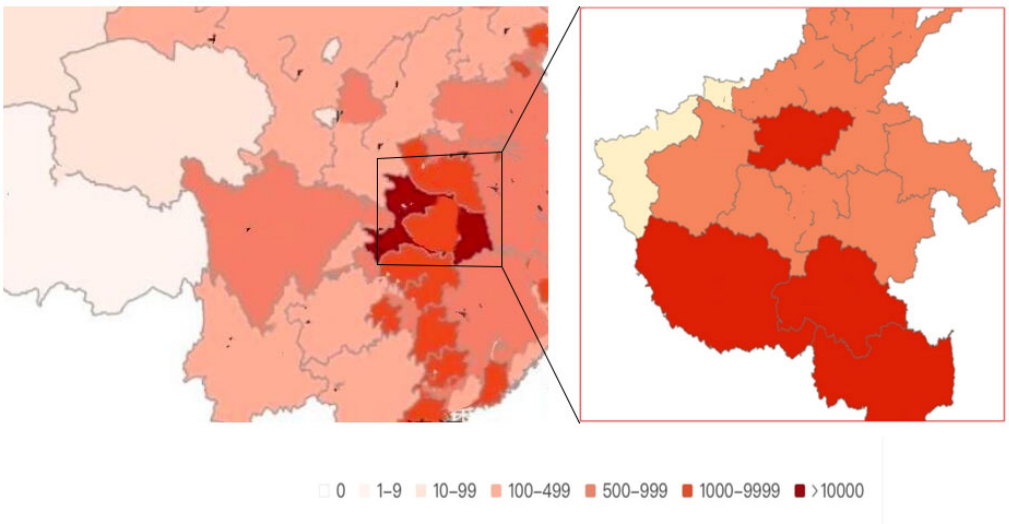

Figure 7. Epidemic distribution map

5.2. Characteristics analysis of flow data

The epidemiological characteristics of epidemiological data were analyzed interactively based on case epidemiological data set combined with map index. In this paper, In this paper, the center coordinate of the map is set to a place, the zoom level was set to 12, scroll zoom was enabled, map style controls, zoom/pan controls and hawk-eye thumbnail controls were added, and all track points were finally displayed as shown in Figure 8.a. Each track point could obtain the basic information of the patient. The flow adjustment data was analyzed from three dimensions: time, space and case information. The time dimension includes the time of case occurrence; The spatial dimension includes

the location of the patient; The case information dimension contains the statistical information of the case, such as the patient's name, gender, epidemic type and status, etc. The distribution of patients can be shown in the heat map as shown in Figure 8.b, where the shaded differences represent the total number of confirmed cases. The darker the shadow, the more severe the outbreak, i.e., the higher the number of confirmed cases; Trace points of the same patient are connected into a line, and the detailed trace text information is presented, as shown in Figure 8.c, to obtain the final visual analysis.

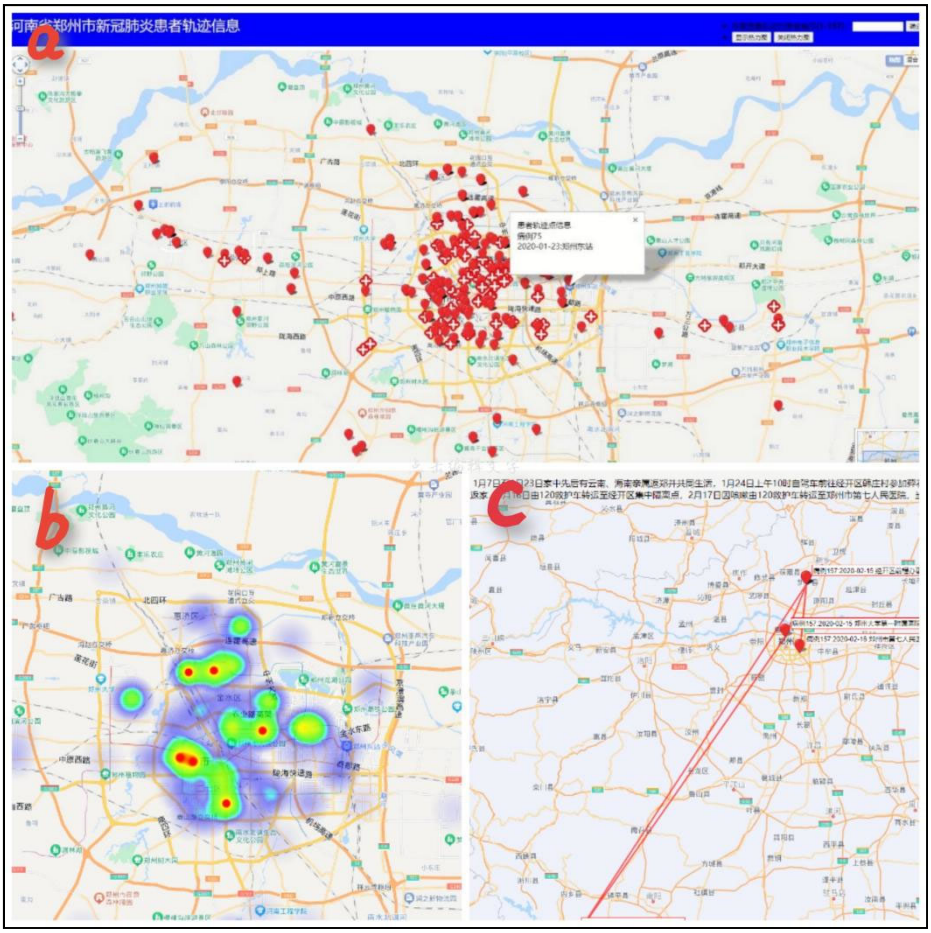

Figure 8. Visual analysis diagram of results

6. CONCLUSION

This article focuses on the epidemiological aspects of the outbreak data. From January 1, 2020 to March 1, 2020, the flow adjustment data of a place was used as the data source, and the data set was theoretically migrated based on the expectation maximization (EM algorithm) statistics to realize the multi-trajectory fusion algorithm, and finally the trajectory fitting of the infected population was carried out. In this paper, the data source is very clear, authoritative and credible, the visual interface is relatively comprehensive, and the analysis process is complete, which can better show the regularity and characteristics of the epidemic flow data, and can be extended to the typical epidemic flow analysis of infectious diseases. However, the visualization system is not perfect enough, in the follow-up research, the visual interface will be more exquisite design and presentation.

ACKNOWLEDGEMENTS

I would like to thank members of the Visual Analysis and Human-Computer Interaction Team of Sichuan University of Science & Engineering for their help.

This research was supported by the Graduate Innovation Fund project of Sichuan University of Science & Engineering, Project Number: Y2022144.

REFERENCES

[1] Yuan Yunxiao, Wang Lining, Wang Baohai, Li Dongyan. Spatial econometric analysis of the concentration and spread of infectious diseases in China: Based on spatial panel data [J]. Mathematics in Practice and Cognition, 2020, 50 (21): 144-150.

[2] Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2[J/OL]. Nat. Microbiol, 2020, 5(4): 536-544.

[3] FEDER K A, PEARLOWITZ M, GOODE A, et al. Linked clusters of SARS-CoV-2 variant B.1.351-Maryland, January-February 2021[J]. MMWR, 2021, 70(17): 627-631.

[4] [4]Du Min, Liu Min, Liu Jue. Epidemiological characteristics and prevention and control progress of novel coronavirus Delta variant [J]. Chinese Journal of Epidemiology, 2021, 42(10): 1774-1779.

[5] [5][Wu B B. Spatial analysis and visualization of COVID-19 epidemic information[J]. Geomatics, Mapping and Spatial Geographic Information, 2021, 44(10): 20-23+28.

[6] [6] Epidemiologic investigation of a case of associated novel coronavirus infection outside a port [J]. Preventive Medicine,2022,34(04):380-384+388.

[7] Ng Chirk Jenn et al. Relationships between cancer pattern, country income and geographical region in Asia.[J]. BMC cancer, 2015, 15(1) : 613.

[8] Sun Yexiang, Lu Jun, Shen Peng, Zhan Siyan, Gao Pei, Zhang Luxia, Chen Kun, He Na, Lin Hongbo, Shui Liming, Li Liming. A new model of disease prevention and control driven by big data in health care [J]. Chinese Journal of Epidemiology, 2021, 42(08): 1325-1329.Ensheng Dong and Hongru Du and Lauren Gardner. An interactive web-based dashboard to track COVID-19 in real time[J].The Lancet Infectious Diseases,2020,20(5):533-534.

[9] Lee J. G.，Han J.，Whang K. Y. Trajectory clustering: a partition- and- group framework[C]//Proceedings of the 2007 ACM SIGMOD international conference on Management of data. 2007:593-604.

[10] Ester M., Kriegel H.P.,Sander J., et al. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]//Proceeding of the Second International Conference on Knowledge Discovery and Data Mining. Portland:AAAI Press, 1996,96(34):226-231.

[11] Li Zhenhui, Lee J. G., Li Xiaolei,etal. Incremental clustering for trajectories[C]l/International Conference on Dat abase Systems for Advanced Applications. Springer, Berlin, Heidelberg, 2010: 32-46.

[12] Yang Shuliang, Bi Shuoben, Nkunzimana A., et al. A spatial clustering method for taxi passenger trajectory [J]. Computer Engineering and Applications, 2018,054(014):249-255.

[13] Ankerst M., Breunig M.M., Kriegel H. P., et al.OPTICS: Ordering points to identify the clustering structure[J]. ACM Sigmod record,1999,28(2):49-60.

[14] Munaga H., Sree M.D.R.M., Murthy J.V.R. DenTrac: A Density based Trajectory Clustering Tool[J]. International Journal of Computer Applications，2012，41(10):17-21.

[15] Zaghlool E.，Elkaffas S., Saad A. A Density-Based Clustering of Spatio-Temporal Data[J].Advances in Intelligent Systems & Computing, 2015,354:41-50.

[16] Hwang J. R.， Kang H. Y.,Li K.J. Spatio-temporal similarity analysis between trajectories on road networks[C]//International Conference on Conceptual Modeling.Springer, Berlin, Heidelberg, 2005: 280-289.

[17] Andrienko G.,Andrienko N. Fuchs G., et al. Clustering trajectories by relevant parts for air traffic analysis[J].IEEE transactions on visualization and computer graphics,2017,24(1): 34-44.

[18] Wang Kan, Mei Kejin, Zhu Jiahui, et al. Hot spot extraction based on spatiotemporal trajectory [J]. Journal of University of Electronic Science and Technology of China,2019,v.48(06): 127-132.

[19] Gong Xi, Pei Tao, Sun Jia, et al. Progress in spatiotemporal trajectory clustering [J]. Progress in Geography, 2011,30(05):522-534.

[20] Cao N, Lin Y R, Sun X, et al. Whisper: Tracing the spatiotemporal process of information diffusion in real time[J]. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(12): 2649-2658.

[21] Schaufler K, Semmler T, Pickard D J, et al. Carriage of extended-spectrum beta-lactamase -plasmids does not reduce fitness but enhances virulence in some strains of pandemic E. coli Lineages[J]. Frontiers in Microbiology, 2016, 7(336): 1-12.

[22] Kurzhals, K., Hlawatsch, M., Heimerl, F., Burch, M., Ertl, T., &Weiskopf, D. Gaze stripes: Image-based visualization of eye tracking data[J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1): 1005-1014.

[23] Guo D. Visual analytics of spatial interaction patterns for pandemic decision support[J]. International Journal of Geographical Information Science, 2007, 21(8): 859-877.

[24] OpenDataCity, Visitor flow analysis by publica wireless[OL].2014-11-06. http://apps.opendatacity.de/relog.

[25] Liu H, Gao Y, Lu L, et al. Visual analysis of route diversity [C]// Proceedings of IEEE Conference on Visual Analytics Science and Technology. Los Alamitos: IEEE Computer Society Press, 2011: 171-180.

[26] Hägerstrand T. What about people in regional science? [J]. Papers in Regional Science, 1970, 24:6-21.

[27] Tominski C, Schumann H, Andrienko G, et al. Stacking-based visualization of trajectory attribute data [J]. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(12): 2565-2574.

[28] Inselberg A. The plane with parallel coordinates [J]. The Visual Computer, 1985, 1(2): 69-91.

[29] Andrienko G, Andrienko N. A general framework for using aggregation in visual exploration of movement data [J]. The Cartographic Journal, 2010, 47(1): 22-40.

[30] Han J, Stefanovic N, Koperski K. Selective materialization: An ecient method for spatial data cube construction [M]. Lecture Notes in Computer Science. Heidelberg: Springer, 1998, 1394: 144-158.

[31] Tobler W. Experiments in migration mapping by computer [J]. Cartography and Geographic Information Science, 1987, 14(2): 155-163.

[32] Gaffney S, Smyth P. Trajectory clustering with mixtures of regression models [C] // Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 1999: 63-72.

[33] Andrienko G, Andrienko N, Heurich M. An event-based conceptual model for context-aware movement analysis [J]. International Journal of Geographical Information Science, 2011, 25(9): 1347-1370.

[34] Dodge S, Weibel R, Lautenschütz A. K. Towards a taxonomy of movement patterns[J]. Information Visualization, 2008, 7(3): 240-252.