Scientific Research Publishing

Journal of Computer and Communications, 2022, 10, 84-94
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Visual Analysis and Prediction System for Infectious Diseases Based on Improved SIR Model

Yu Qiu, Yadong Wu*, Qibiao Wang, Weihan Zhang

School of Computer Science and Engineering, Sichuan University of Science & Engineering, Yibing, China
Email: *wyd028@163.com

How to cite this paper: Qiu, Y., Wu, Y.D., Wang, Q.B. and Zhang, W.H. (2022) A Visual Analysis and Prediction System for Infectious Diseases Based on Improved SIR Model. *Journal of Computer and Communications*, **10**, 84-94.
https://doi.org/10.4236/jcc.2022.1012007

Received: November 4, 2022
Accepted: December 25, 2022
Published: December 28, 2022

Abstract

To effectively track the impact of population migration between regions on the spread of infectious diseases, this paper proposes a visualized analysis and prediction system of infectious diseases based on the improved SIR model. The research contents including: using the multi graph link interaction mode, visualizing the space-time distribution and development trend of infectious diseases; The LightGBM model is used to track the changes of infection rate and recovery rate, and the Mi/Mo SIR model is constructed according to the initial data of different populations; Mi/Mo SIR model is used to predict infectious diseases in combination with visual panel, providing users with tools to analyze and explain the space-time characteristics and potential laws of infectious diseases. The study found that the closure of cities and the restriction of personnel mobility were necessary and effective, and the system provided an important basis for the prediction and early warning of infectious diseases.

Keywords

Migration, Epidemic Forecast, Mi/Mo-SIR Model, Machine Learning, Visual Analytics System

1. Introduction

In recent years, not only various new infectious diseases have emerged frequently, but also many familiar infectious diseases in the past are still helpless and have brought serious impacts on people's lives, such as atypical pneumonia (SARS), H1N1, and Ebola virus. Since December 2019, COVID-19 (Corona Virus Disease 2019) has been characterized by rapid spread and high social panic, with an exponential increase in the cumulative number of illnesses, gradually developing into a major problem affecting global health and economy. In order to control the spread of the epidemic, there is an urgent need to contain the spread of infectious diseases in different regions. Due to the specific nature of infectious diseases, the process of their spread cannot be reproduced by experimental simulations, making it necessary to analyze this process using mathematical modeling. Infectious disease transmission kinetic modeling is an important method for theoretical and quantitative study of the epidemiological patterns of infectious diseases, which is based on mathematical models of the transmission kinetic properties of infectious diseases based on the growth characteristics of the population, the pattern of disease transmission within the population, and the social factors associated with it [1]. The SIR (Susceptible-Infectious-Recovered) model, as the most representative infectious disease transmission kinetic model, has made a seminal contribution to the study of infectious diseases.

In the study of infectious diseases, effective means are often needed to mine and analyze the large-scale data therein. Data visualization is a method based on information graphics, image processing, and statistical graphics, and uses interaction theory to transform data into a graphical display [2]. By transforming abstract and static epidemic data into visual and dynamic interactive graphs through visualization means, users can be effectively guided to explore the hidden valuable information behind the epidemic data.

In this context, this paper applies and investigates an improved SIR model to the prediction of COVID-19, introduces the population migration (Move-in/ Move-out) factor into the traditional SIR model, and proposes an improved Mi/Mo-SIR (Moved-in/Moved-out-Susceptible-Infectious-Recovered) model, using the machine learning LightGBM framework [3] for conducting the model for infectious rate and recovery rate tracking, using a two-level interactive spatial-temporal visualization system, and providing a multi-graph linked interaction model to provide users with visualization of spatial-temporal distribution of the epidemic, trend prediction, etc. The interactive analysis helps users to easily analyze and interpret the spatial-temporal characteristics and potential patterns of epidemics, and provides an important basis for epidemic prediction and early warning.

2. Literature Review

In 1926, Kermack and McKendrick [4] [5] proposed the famous SIR model using a kinetic approach became a widely recognized and studied model of infectious disease transmission kinetics. Following the outbreak, related research fields and researchers have conducted a lot of studies and applications of the SIR model for the development process and transmission trend of the epidemic.

Lee *et al.* [6] proposed a SIR model with human intervention factors for the isolation intensity of prevention and control policy; Liu *et al.* [7] proposed a class of SIR models with regional switching to study the potential of virus transmission in two populations; Araiinejad *et al.* [8] developed a time-dependent SIR model with parameter estimation using the Lasso regression [9] algorithm to monitor the intervention effectiveness; Goel *et al.* [10] proposed a population mobility-based SIR model that considered the real population distribution in different regions of the world and connectivity factors between regions of the world and found that limiting mobility in the top 10% of connected locations could reduce the number of infections by 18% to 27%; Deo *et al.* [11] explored the impact of undetected cases on the spread of infectious diseases and proposed a dynamic SIR model SI(Q/F)RD (Susceptible-Infected-(Quarantined/Free)-Recovered-Deceased).

In the fight against COVID-19, large-scale epidemic data, such as case data, spatial-temporal data, etc., were generated, and the massive amount of complex data posed a huge challenge to understanding, requiring effective means to mine and analyze the large-scale data. The widespread use of data journalism has greatly enhanced the audience's intuitive understanding of the epidemic form [12], and data visualization techniques and infectious disease transmission dynamics models are of great relevance to explore the transmission process of COVID-19, predict epidemic trends and make prevention and control decisions.

Since to the current work on SIR model in epidemic prediction has not taken into account the influence of inter-regional population migration on the spread of infectious diseases, this paper takes the current as well as future needs of infectious disease prevention and control as the starting point, uses the relationship between urban migration index and epidemic spread for visualization research, uses Mi/Mo-SIR model for quantitative analysis, dynamic prediction and evaluation of the epidemic development status, and through the LightGBM method tracks the changes of transmission rate and recovery rate over time, provides spatial-temporal visual analysis of the transmission process of infectious diseases, recognizes the inner law and predicts the development process of diseases, and helps users to make better prevention and control decisions.

3. Methodology

3.1. Overall System Architecture

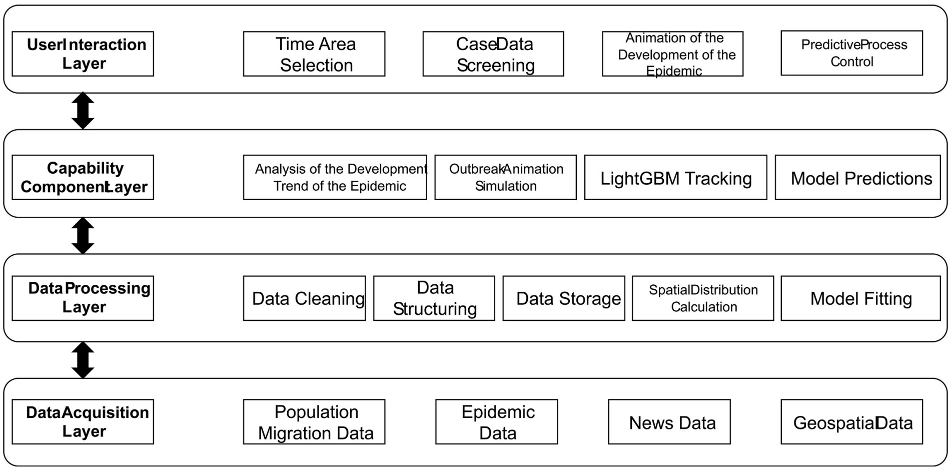
The system contains two main sections: visual analysis of the temporal and spatial situation of infectious diseases, and visualization of infectious disease trend prediction. The overall architecture of the system is shown in **Figure 1**.

Taking COVID-19 as an example, an automated collection method was used to crawl the publicly available data on the web between January 22 and December 12, 2020 through the Python language, and the data from each platform was integrated using manual, and the accuracy of the data was cross-validated through multi-source data collection [13]. The crawled data were transformed into usable Json structured data after data cleaning to facilitate data storage and data analysis. The statistical case data and geospatial data were used to analyze the spatial-temporal development trend of the epidemic using visualization methods, and the Mi/Mo-SIR model was used to predict the propagation process to the epidemic. At the same time, it allows users to select different time spans in the visualization interactive panel to filter the case data and control the spatial-temporal progression of the prediction.

3.2. Data Acquisition and Processing

Data collection is performed using the Scrapy-Redis method, an open source framework for Python. Scrapy-Redis is a component of the Scrapy framework based on the Redis database for distributed development and deployment of Scrapy projects, and the overall operational flow is shown in **Figure 2**.

The data processing process mainly includes: data cleaning, data structured organization, data storage, patient spatial distribution calculation, etc. The data processing is completed by automated technical means and manual cleaning by personnel, and the data processing flow chart is shown in **Figure 3**.



UserInteraction Layer
Time Area Selection
CaseData Screening
Animation of the Development of the Epidemic
PredictiveProcess Control
Capability ComponentLayer
Analysis of the Development Trend of the Epidemic
OutbreakAnimation Simulation
LightGBM Tracking
Model Predictions
DataProcessing Layer
Data Cleaning
Data Structuring
Data Storage
SpatialDistribution Calculation
Model Fitting
DataAcquisition Layer
Population Migration Data
Epidemic Data
News Data
GeospatiaData

Figure 1. System architecture diagram.

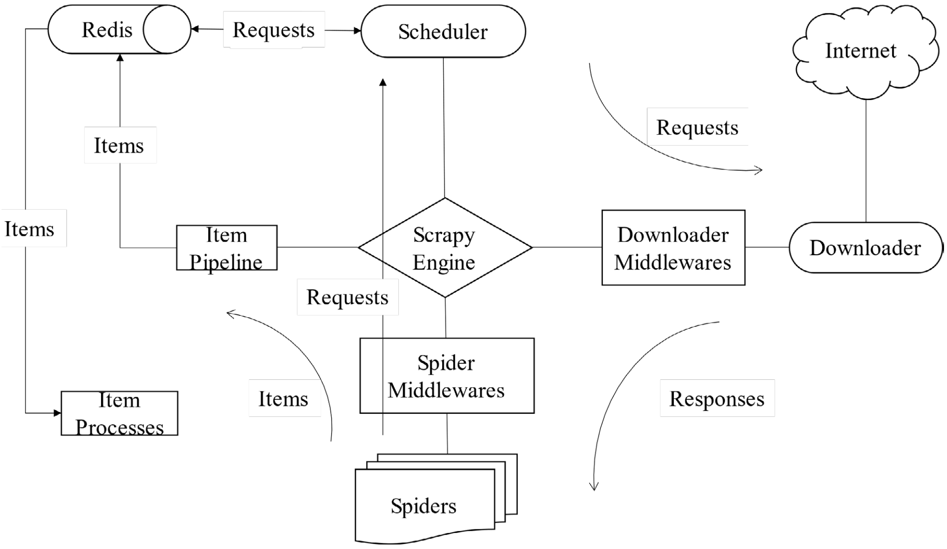


Figure 2. Scrapy-Redis crawler framework.

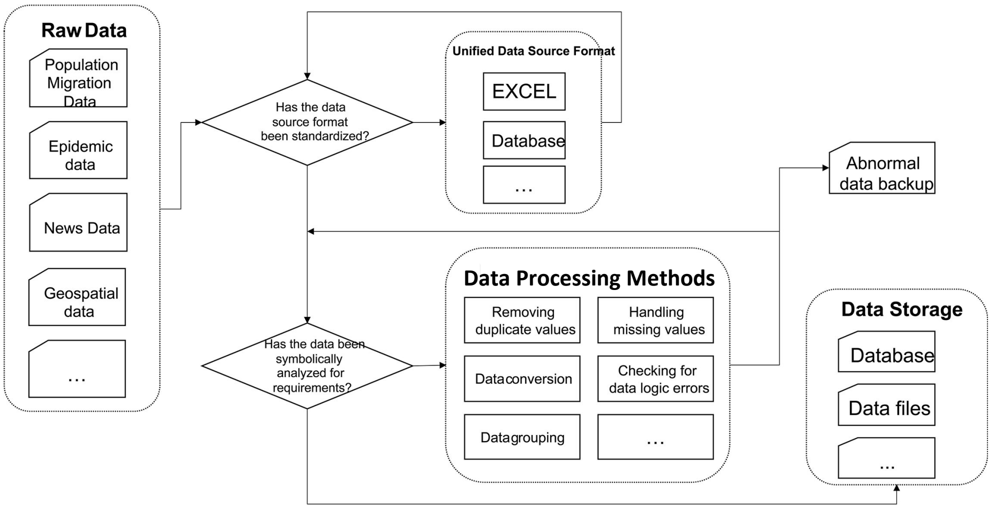

Raw Data
Population Migration Data
Epidemic data
News Data
Geospatial data
…
Has the data source format been standardized?
Unified Data Source Format
EXCEL
Database
…
Abnormal data backup
Has the data been symbolically analyzed for requirements?
Data Processing Methods
Removing duplicate values
Handling missing values
Dataconversion
Checking for data logic errors
Datagrouping
…
Data Storage
Database
Data files
…

Figure 3. Data processing flow chart.

The processed population migration data, epidemic data, news data, geospatial data, etc. are stored in CSV file format with text for subsequent analysis and prediction.

3.3. Mathematical Modeling and Optimization

The classical SIR equation assumes a constant size of the susceptible population, but in the real situation, the population is dynamic and the population moving in and out of the city can lead to the movement of infected people. Therefore, the improved Mi/Mo-SIR model in this study introduces the dynamic changes of the susceptible population (*S*) and the infected population (*I*) by moving into the *Mi*(*t*) bin and moving out of the *Mo*(*t*) bin to simulate the dynamic changes of the population, and the transformation relationship of the population is shown in **Figure 4**.

The base model is as follows expression (1).

$$\begin{cases} \dfrac{\mathrm{d}S(t)}{\mathrm{d}t} = \dfrac{\beta(t)S(t)I(t)}{N(t)} \\ \dfrac{\mathrm{d}I(t)}{\mathrm{d}t} = \dfrac{\beta(t)S(t)I(t)}{N(t)} - \gamma(t)I(t) \\ \dfrac{\mathrm{d}R(t)}{\mathrm{d}t} = \gamma(t)I(t) \end{cases} \tag{1}$$

The research modified model is given by expression (2).

$$\begin{cases} \dfrac{\mathrm{d}S(t)}{\mathrm{d}t} = S_{in}(t) - \dfrac{\beta(t)S(t)I(t)}{N(t)} - S_{out}(t) \\ \dfrac{\mathrm{d}I(t)}{\mathrm{d}t} = \dfrac{\beta(t)S(t)I(t)}{N(t)} + I_{in}(t) - I_{out}(t) - \gamma(t)I(t) \\ \dfrac{\mathrm{d}R(t)}{\mathrm{d}t} = \gamma(t)I(t) \end{cases} \tag{2}$$

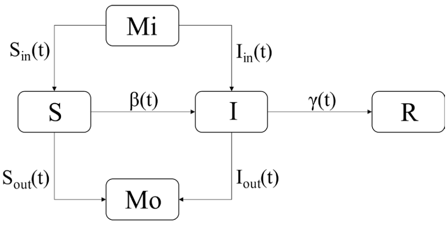


Figure 4. Mi/Mo-SIR model state transfer diagram.

where,

$$\begin{cases} S_{in}[t] = Mi[t] \times (1 - P_{in}[t]) \\ S_{out}[t] = Mo[t] \times (1 - P_{in}[t]) \\ I_{in}[t] = Mi[t] \times P_{in}[t] \\ I_{out}[t] = Mo[t] \times P_{in}[t] \\ N[t+1] = N[t] + Mi[t] - Mo[t] \\ P_{in}[t] = \dfrac{S[t+1] - S[t]}{N[t+1]} \end{cases} \tag{3}$$

The Mi/Mo-SIR model parameters are explained as follows.

$N(t)$: the total population of the province.

$S(t)$: represents susceptible, a group of people who may become infected after contact with infected persons.

$I(t)$: represents infectious, a group of people who are infectious.

$R(t)$: represents the number of people who recovered or died after infection.

$M(t)$: represents moved-in, a population of people who have moved into the current area from other areas.

$O(t)$: represents moved-out, the population of people who moved into the current area from other areas.

S_{in} / S_{out}: the number of people moving in and out of the susceptible population.

I_{in} / I_{out}: the number of people moving in and out of the infected population.

$P_{in}(t)$: the probability of infected persons in the moved-in population.

$\beta(t)$: probability of transmission from a susceptible population to an infected population.

$\gamma(t)$: removal rate of the disease (including cure rate and mortality rate).

4. Research Results

4.1. Dataset

The research uses case data including daily new confirmed cases, existing cases, cumulative confirmed cases, cured cases, and deaths, and migration data including daily in-migration and out-migration size indices. Considering that there is a serious disconnect between confirmed cases in Hubei Province and other provinces and cities, and that most cities in Hubei Province adopted the "city closure" from January 23, 2020, data from Hubei Province were excluded from this study. Table 1 shows the top 10 provinces (except Hubei Province) with the number of confirmed COVID-19 cases from January 22, 2020 to December 12, 2020.

4.2. Tracking Infection and Recovery Rates Using the LightGBM Model

LightGBM is an efficient and open source distributed gradient framework based on the GBDT (Gradient Boosting Decision Tree) algorithm [14] released by Microsoft. The LightGBM framework proposes two new methods, GOSS (Gradient-based One-Side Sampling) and EFB (Exclusive Feature Bundling) to accelerate the training process of the model by calculating the information gain of some samples. The framework supports feature parallelism, data parallelism and voting parallelism strategies to reduce the time complexity while improving the accuracy, which is a good solution to the processing of massive data and is widely used for regression and classification prediction.

Figure 5 and Figure 6 depict the prediction of transmission and recovery rates in China (excluding Hubei Province) from January 22, 2020 to December 12, 2020 using the LightGBM model, which was found to be a good fit for tracking transmission rates by comparing the validation set with the predicted data.

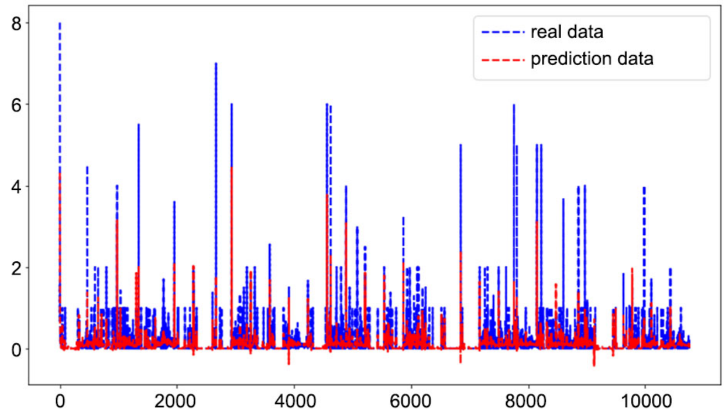


Figure 5. Beta data over time.

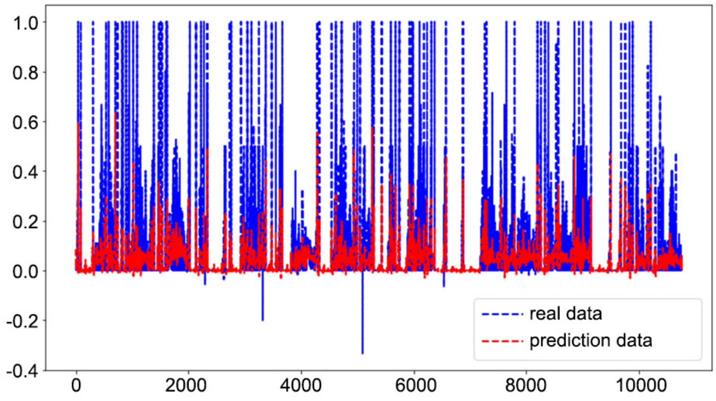


Figure 6. Gamma data over time.

Table 1. Top 10 provinces with confirmed cases as of Dec. 12.

Province	Infected	Recovered	Death
Hong Kong	7446	6114	115
Guangdong	2016	1972	8
Shanghai	1405	1309	7
Zhejiang	1296	1290	1
Henan	1295	1266	22
Hunan	1020	1016	4
Anhui	992	986	6
Xinjiang	980	977	3
Heilongjiang	956	936	13
Beijing	954	940	9

4.3. Visual Analysis of the Spatial and Temporal Situation of the Epidemic

Use the Time Progression Chart to plot the patient situation in all cities and states. Users can drag the timeline to change the time period presented on the map, and click the Play button to watch the changes in the epidemic situation in the whole Sichuan Province; use the New Trend Chart to present the daily new confirmed, cured, and death data, and users can click the confirmed, cured, and death buttons to view the changes in a specific item. The user can click on the Confirmation, Cure, Death button to see the change of a specific item. To view the daily trend of new cases in a particular city, you can click on the corresponding city on the map; use the regional ranking chart to display the confirmed cases in each region, and users can choose to rank them in ascending or descending order. The visualization interface of the spatial and temporal situation of infectious diseases is shown in Figure 7.

Area A shows the regional ranking information, with two bars of different color scales indicating the proportion of local development period and foreign input period; Area B presents the city, existing confirmed, cumulative confirmed, and cumulative factors by parallel coordinates, and users can click the buttons on the map to display different cumulative data; Area C uses a curve to indicate the daily confirmed, cured, and death data. If you want to see the trend of new cases in a certain province or city, you can click the corresponding province or city on the map in area B; area D uses a scatter plot to compare the number of confirmed cases with the number of cured cases; area E allows you to change the time period presented by dragging the time axis, and you can click the play button to view the whole map. As well as the ability to click the play button to view the entire epidemic as it changes.

4.4. Visualization of Outbreak Trend Prediction

The Mi/Mo-SIR model is used to simulate the incidence of an outbreak in a region and the cumulative number of incidences after the outbreak, presenting the transmission of an area through an epidemic spread map, and supporting the risk analysis of associated cities. The system provides default parameters, and users can adjust the values of each parameter by themselves to plot the real-time epidemic trend and the development of risk in the associated areas. The visualization interface for infectious disease trend prediction is shown in Figure 8.

Area A is the parameter control panel for Mi/Mo SIR model simulation. Its default parameters are calculated from real data. Users can adjust the required parameters according to needs to control the start and end of the simulation process; Area B shows the infection status of the associated areas after the outbreak of the epidemic in the simulation area; The C area uses a line chart to represent the daily newly added diagnosis and cure data in this area in the simulation. Users can click the diagnosis and cure buttons above to view the specific changes of an item separately; Area D shows the associated risk level between the region and the six cities and states with the highest migration activity in the simulation process.

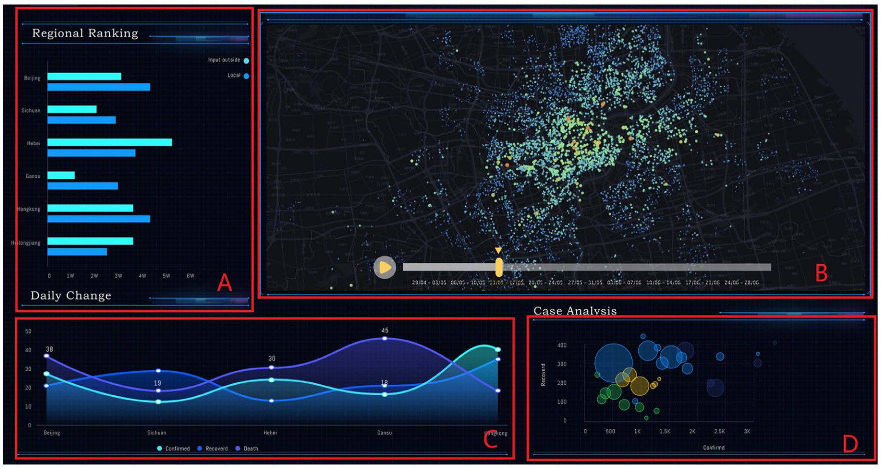


Figure 7. Spatial and temporal infectious disease situation visualization interface.

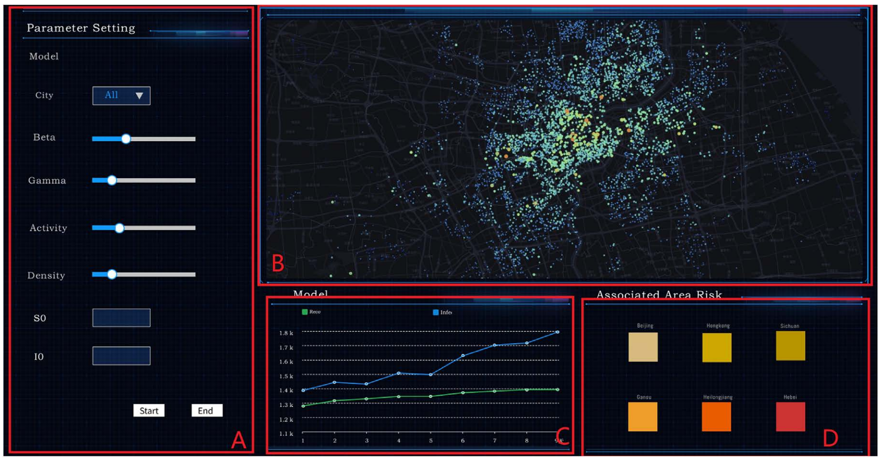


Figure 8. Visualization interface for epidemic trend prediction.

5. Conclusions

This paper considers the spread of epidemics brought by population migration to the region and establishes a visual analysis and prediction system for infectious diseases based on an improved SIR model. Among them, the LightGBM model is used to track the changes of transmission and recovery rates, and a time-dependent population migration Mi/Mo-SIR model is established to make the differential dynamics equation modeling of the epidemic development more realistic. Through the visualization interaction strategy, the daily new diagnosis, cure, and death data are analyzed using a multi-graph linked time series, and combined with map heat maps, the temporal linkage is performed on all graphs by selecting the event time to show different event phenomena occurring at the same time, which assists users in reviewing the effectiveness of policies through prevention and control effects and making reasonable administrative decisions. However, influenced by age, gender, policy changes, etc., the model can hardly avoid some differences with reality, and needs to adjust the epidemic prevention and control strategy according to the actual situation.

Influenced by advanced modern transportation technology, the speed of population movement has accelerated, and different countries and regions of the world have become more interconnected and dependent on each other. The spread of epidemics not only affects neighboring regions, but is also highly susceptible to the formation of large-scale outbreaks in areas with advanced economic and trade development and frequent population movement, and epidemic prevention measures should be taken as early as possible. At the same time, it is necessary to pay attention to imported cases from high-risk areas outside of China and strengthen the prevention and control of international logistics. Vaccination progress also needs to be further advanced at home and abroad to accelerate the construction of epidemic prevention barriers. The next step of this paper will consider the impact of internal and external mobility factors and vaccination intensity on the prevention and control of the epidemic, and provide a basis for decision making and new ideas for the "post-epidemic era".

Fund

School of Computer Science and Engineering Talent Introduction Project (2020RC20).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Xu, R., Tian, X.H. and Gan, Q.T. (2019) Modeling and Analysis of Infectious Disease Dynamics. Science Press, Beijing.

[2] Chen, W., Shen, Z.Q., Tao, Y.B., *et al.* (2015) Data Visualization. Electronic Indus-

try Press, Beijing.

[3] Ke, G., Meng, Q., Finley, T., Wang, T., *et al.* (2017) Lightgbm: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, **30**, 3249-3157.

[4] Thomas, S.A. (2007) Lies, Damn Lies, and Rumors: An Analysis of Collective Efficacy, Rumors, and Fear in the Wake of Katrina. *Sociological Spectrum*, **27**, 679-703. https://doi.org/10.1080/02732170701534200

[5] Zhao, L., Cui, H., Qiu, X., Wang, X. and Wang, J. (2013) SIR Rumor Spreading Model in the New Media Age. *Physica A*: *Statistical Mechanics and Its Applications*, **392**, 995-1003. https://doi.org/10.1016/j.physa.2012.09.030

[6] Lee, Z.L., Zide, W.A.N.G. and Haotian, Y.A.N.G. (2020) Quantitative Factors and Mathematical Modeling of COVID-19 Pandemic under Human Interventions. In 2020 *International Conference on Public Health and Data Science* (*ICPHDS*), Guangzhou, 20-22 November 2020, 257-264.

[7] Liu, L., Jiang, D. and Hayat, T. (2021) Dynamics of an SIR Epidemic Model with Varying Population Sizes and Regime Switching in a Two Patch Setting. *Physica A*: *Statistical Mechanics and its Applications*, **574**, Article ID: 125992. https://doi.org/10.1016/j.physa.2021.125992

[8] Araiinejad, L.S., Carlton, J.R., Li, Y. and Zheng, J. (2020) Assessing the Impact of Government Interventions on the Spread of COVID-19 with Dynamic Epidemic Models: A case study of Texas. *In* 2020 *IEEE International Conference on Bioinformatics and Biomedicine* (*BIBM*), Seoul, 16-19 December 2020, 2274-2281. https://doi.org/10.1109/BIBM49941.2020.9313591

[9] Park, T. and Casella, G. (2008) The Bayesian Lasso. *Journal of the American Statistical Association*, **103**, 681-686. https://doi.org/10.1198/016214508000000337

[10] Deo, V. and Grover, G. (2021) A New Extension of State-Space SIR Model to Account for Underreporting—An Application to the COVID-19 Transmission in California and Florida. *Results in Physics*, **24**, Article ID: 104182. https://doi.org/10.1016/j.rinp.2021.104182

[11] Goel, R. and Sharma, R. (2020) Mobility Based Sir Model for Pandemics-With Case Study of Covid-19. In 2020 *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (*ASONAM*), The Hague, 7-10 December 2020, 110-117. https://doi.org/10.1109/ASONAM49781.2020.9381457

[12] Xu, X.D. (2017) Data News. China Renmin University Press, Beijing.

[13] Dong, E., Du, H. and Gardner, L. (2020) An Interactive Web-Based Dashboard to Track COVID-19 in Real Time. *The Lancet Infectious Diseases*, **20**, 533-534. https://doi.org/10.1016/S1473-3099(20)30120-1

[14] Qiu, X., Zhang, R., Xu, H. and Li, X. (2021) Local Interpretable Explanations for GBDT. In 2021 *International Joint Conference on Neural Networks* (*IJCNN*), Shenzhen, 18-22 July 2021, 1-10.