Mining and forecasting of infectious disease transmission data based on smart cities

Jiongxu Mou*
School of Management, Wuhan University of Technology, China

*Corresponding author:463730589@qq.com

Abstract. With the rise of a new round of urbanization and the growing public awareness of public health, the traditional infectious disease management system can no longer meet the health management needs of smart cities. Establishing statistical indicators and combined time series analysis and prediction models that reasonably describe the spread of infectious diseases can explore the basic trends of infectious disease transmission and make reasonable predictions. In this paper, the number of patients and deaths in the provinces and cities of tuberculosis from 2004 to 2016 were used. The data of tuberculosis from 2017 to 2019 was predicted by the ARMA model. The model was tested by comparison to prove that the method of mining and prediction was reasonable. The health management department of the smart city formulates policies to provide reference.

1. Introduction
Since the beginning of the 21st century, the continuous development of China's urbanization has brought about economic growth, but at the same time there have been many problems, such as excessive urban population density, traffic congestion, and serious environmental pollution. There is a need to introduce a more intelligent approach to managing cities[1]. At present, domestic and foreign scholars' research on smart cities mainly focuses on the elaboration of connotation concepts, lacking the determination and specific application of indicator data. Regarding the concept of the connotation of smart cities, Li Deren[2] and others regard smart cities as urban information management systems under the comprehensive digitalization of cities and the application of Internet of Things technologies; Jin Jiangjun[3] and others regard smart cities as Internet of Things, cloud computing, artificial intelligence, Data mining and other technical means to improve the city's service efficiency. Therefore, I believe that smart cities are based on information technology and can establish a practical, powerful and perfect application system for public management, public services and public industries to improve the efficiency of government services and improve people's quality of life.
The main application areas of smart cities are concentrated in urban public security, urban infrastructure, urban transportation, energy and public utilities, environmental protection, and smart governance. In the field of urban public safety, the control and prevention of infectious diseases is particularly important[4]. The government's health department should prevent or respond to all kinds of sudden infectious diseases in a timely and effective manner. Therefore, it is necessary to establish a sensitive and effective early warning system, collect epidemic data and make countermeasures. At present, there are few studies at home and abroad in the field of infectious disease control and prevention in smart cities, which is worthy of further exploration.

The research content of this paper belongs to the category of data mining and modeling prediction in the field of public security in smart cities. This article takes the spread of tuberculosis as a case, collects information data, focuses on the occurrence of tuberculosis infection, possible influencing factors, and builds a combined forecasting model based on past data and mining results to spread the tuberculosis disease in the future. Forecasting, I hope to take tuberculosis as an example to achieve timely warning before the large-scale outbreak of infectious diseases[5].

2. Research methods

2.1. Selection of data and establishment of indicators:

We take tuberculosis as an example. Tuberculosis is an ancient infectious disease, and countries all over the world have been fighting tuberculosis. Although scientists have made many achievements in diagnostic technology, drug development, and treatment for many years, tuberculosis is still one of the top ten causes of death, second only to AIDS. At present, the important challenges of tuberculosis prevention and treatment in China are low tuberculosis detection rate, low pathogen diagnosis rate, low screening rate and high drug resistance rate.

We selected the number of cases and occupational distribution of tuberculosis in each province and municipality from 2006 to 2016, and wanted to predict the epidemiological trend of tuberculosis and predict the number of cases and the number of deaths in 2019. This paper selects early warning indicators - morbidity and mortality to determine the prevalence of tuberculosis and the extent of harm[6].

2.1.1. Incidence rate

The incidence rate indicates the frequency of occurrence of new cases of a disease in a certain population within a certain period of time. The incidence rate can more fully reflect the prevalence of infectious diseases. The formula is (the number of people infected with infectious diseases per 100,000 people): 

$$\text{Incidence rate} = \frac{\text{new diseases in a certain population within a certain period of time}}{\text{Simultaneous population}} \times k$$

As one of the core statistical indicators of medicine, the incidence rate can effectively judge the epidemic of infectious diseases. If the incidence of a certain infectious disease in a certain region or unit exceeds the level of the same period of the previous year, it can be considered that there is a trend of infectious diseases. It is conducive to describing the distribution of disease, proposing the cause hypothesis, and evaluating the effects of prevention and treatment measures.
We calculated the results of the 2004-2016 incidence as follows (number of cases per 100,000 people):

| year | 2004 | 2005 | 2006 | …… | 2014 | 2015 | 2016 |
|------|------|------|------|-----|------|------|------|
| Incidence | 74.64 | 296.40 | 85.78 | …… | 65.02 | 62.85 | 60.45 |

2.1.2. Mortality rate

The case fatality rate indicates the proportion of all patients suffering from a disease due to the disease within a certain period of time, and is used to describe the severity of a particular disease. Its formula is:

\[
\text{Mortality rate} = \frac{\text{Number of deaths due to illness in a period}}{\text{Number of patients in the same period}} \times 100\%
\]

As a medical statistical indicator, mortality is huge. Diseases with high mortality rates, such as rabies and AIDS, have long been the focus of prevention and medical challenges. When the mortality rate rises compared with the same period of the previous year, it should be highly valued by the relevant departments. It can be seen that the case fatality rate can reflect the severity of the disease and reflect the medical ability ability such as treatment ability from the side, and is also an important indicator for evaluating infectious diseases.

We calculated the partial outcomes for the 2004-2016 mortality rate as follows (number of deaths per 100 patients due to illness):

| year | 2004 | 2005 | 2006 | …… | 2014 | 2015 | 2016 |
|------|------|------|------|-----|------|------|------|
| Mortality | 0.148 | 0.270 | 0.296 | …… | 0.252 | 0.264 | 0.295 |

2.2. Data basic trend analysis:

![Patient number histogram](image)
From 2004 to 2016, the tuberculosis infectious disease reported 9663729 cases of the infectious disease, with an average annual incidence rate of 81.45/100,000. The incidence of the disease showed a slow downward trend. During the period, there were two growth peaks, which were in 2005 and 2008 respectively, and the abnormal growth peak appeared in 2005.

![Fig2. Death toll histogram](image)

From 2004 to 2016, the number of deaths caused by tuberculosis reached 36,493, and the average mortality rate was 2,802.75 person/year. The number of deaths declined from 2009, but there was an abnormal increase from 2005 to 2009.

### 2.3 The establishment of the prediction model:

By analyzing the time series of the number of cases and the number of deaths caused by tuberculosis, it is believed that the number of cases and the number of deaths caused by infectious diseases contain both deterministic dynamic trends and non-stationary sequences containing random fluctuations. Therefore, we want to predict the number of cases and the number of deaths caused by the ARIMA model for the country's infectious diseases[7].

#### 2.3.1. Time series processing and stationarity analysis

From the original sequence diagram, as shown below, there is a trend in the number of infectious diseases in 2004-2016, and data preprocessing includes sequence smoothing and non-randomization[8].

The purpose of the smoothing is to reduce the number of random variables by the sequence and increase the sample size of the variable to be estimated. That is to simplify the difficulty of timing analysis and improve the estimation accuracy of the mean function. The fitting curve obtained from the sample time series can continue to "inertial" along the existing form for a period of time in the future[9].

We make a first-order difference to the original sequence: calculate the difference between the time series and the time, obtain the sequence diagram of the first-order difference, and test the stationarity.

Use the adftest() function in MATLAB to perform ADF test (refer to whether there is a unit root in the test sequence, because the unit root is a non-stationary time sequence. It can be proved that the unit root process is not stable in the sequence) and the return value is 1, Explain that the time series after the difference is stable. It is convenient to further judge the model and solve.

### 2.3.2. Model identification and ordering
Model identification for time series can be determined by its autocorrelation function and partial autocorrelation function. The area between the solid blue lines in the above figure is within the 95% confidence interval of the autocorrelation or partial autocorrelation that is positive and negative 2 times the estimated standard deviation. By observing the picture, we can choose the autocorrelation function and the order of the partial auto-correlation is not zero, and determine the multiple ARIMA models. Among them, the ARIMA (1,1,2) model has the smallest AIC value and SC value, and the model is determined by AIC and SC criterion test: ARIMA (1, 1, 2) model[10].

2.3.3 ARIMA model demand forecasting

Using the ARIMA model to predict the number of deaths in 2004-2019, it can be seen from the figure that the prediction results are more accurate. By calculating the relative error, it is found that the relative error of most years is within 5%, achieving high precision, but there is still room for improvement.

2.4 residual white noise test

White noise is a series of independently distributed normal series. The sequence is irregular and can be repeated and oscillated at the mean without affecting the trend of the model. White noise test is performed on the residual, LB(6)=7.62, P>0.05, indicating that the residual is a white noise sequence. It is suggested that the model is valid[11].

3. conclusion and suggestion

The spread of infectious diseases affects people's normal life or property safety to varying degrees. This paper focuses on the occurrence and possible influencing factors of infectious diseases in cities, and establishes prediction and early warning models based on mining results. On the one hand, it can predict the number of patients and the number of deaths most likely to occur in the future transmission of infectious diseases, and prevent and control possible outbreaks of infectious diseases in advance. On the other hand, it can monitor and warn the spread and evolution of infectious diseases, so it can achieve effective information sharing and coordination after the outbreak of infectious diseases.

In order to cope with the large-scale outbreak of infectious diseases, our proposal is based on the data of the health management department of the smart city, aiming at effectively grasping the methods of data collection, analysis and prediction from the perspective of the epidemic trend of infectious diseases and making decisions. Government departments should organize corresponding
law enforcement activities in conjunction with the seasonal characteristics of infectious diseases, and strengthen supervision and inspection. For example, winter and spring seasons should focus on respiratory infectious diseases, and summer and autumn should focus on the prevention and management of intestinal infectious diseases.

References
[1] DEAKIN M. Smart cities: governing, modeling, and analyzing the transition[M]. Oxford: Routledge, 2013.
[2] Li Deren, Shao Zhenfeng, Yang Xiaomin. Theory and Practice from Digital City to Smart City. [J].Geospatial Information 2011, 9(06): 1-5.(in Chinese)
[3] Jin Jiangjun. Smart government: a new stage of e-government development. [J].Informatization construction 2011(11): 16-17.(in Chinese)
[4] PIERRE J. Models of urban governance: the institutional dimension of urban politics[J]. Urban Affairs Review, 1999, 34(3): 372-396.
[5] Zeng Jing, Li Rong, Gu Changmei, Nie Shaofa. Application of Delphi Method to Construct Early Warning Index System of Infectious Diseases in Wuhan City Circle.[J].Chinese Journal of Social Medicine 2012, 29(03): 211-213.(in Chinese)
[6] He Wei, Research on Public Security Risk Management and Control of Smart Cities Based on Big Data.[J].In 2019 National Public Safety Communication Symposium, Urumqi, Xinjiang, China (2019); p5(in Chinese)
[7] Du Longfei, Tian Zhaojun, Lu Yi, Yin Yafei. Analysis of the current situation of public security emergency management in smart cities under the era of big data and countermeasures.[J]. Security 2018, 39(11): 50-52.(in Chinese)
[8] Contreras J, Espinola R, Nogales F J, et al. ARIMA models to predict next-day electricity prices[J]. IEEE Transactions on Power Systems, 2003,18(3):1014-1020
[9] Boroojeni K G, Amini M H, Bahrami S, et al. A novel multi-time-scale modeling for electric power demand forecasting: From short-term to medium-term horizon[J]. Electric Power Systems Research, 2017,142:58-73.
[10] Li Peng. Research on the application of infectious disease prediction based on time series [D]: Kunming University of Science and Technology, 2018.(in Chinese)
[11] Wang Yu, Liu Guangwen, Jia Lei. Time series analysis of epidemiological characteristics of intestinal infectious diseases in Hotan City from 2010 to 2014.[J].China Public Health 2016, 32(09): 1265-1267.(in Chinese)