**Analysis of 2019-nCoV prevention effect based on immigration data in China**

Qiangsheng Huang

1. **Abstract**

In this paper, in response to the outbreak of novel coronavirus (2019-nCoV)–infected pneumonia (NCIP) in China in early 2020, a model was established based on the daily immigration population data of provinces / municipalities in China and the daily number of newly diagnosed patients in each region. Offset and time window, which are the model parameters used for the best fit between the daily risk coefficient and daily number of newly diagnosed patients, is used as the effect assessment of the external and internal migration control. The results show that the prevention and control of Fujian and some other regions is relatively good, but Heilongjiang and some other regions are not good. The accuracy of the model has been confirmed from many aspects such as theory, data, and news.

Key words: novel coronavirus (2019-nCoV)–infected pneumonia (NCIP); Epidemic Prevention and Control; Big Data Analysis; Data mining; Mathematical modeling

1. **Background**

It has been 2 months since the outbreak of novel coronavirus (2019-nCoV)–infected pneumonia (NCIP) in Wuhan, Hubei, since December 2019. Chinese has been ignorant of it from the beginning, and then gradually paid attention to the investigation and control of incoming immigration population, and closed down some densely populated regions. And minimize internal people going out, wearing masks when going out, community prevention, and so on. At present, the epidemic in the regions other than Hubei, which is the region where the virus originated, have been basically controlled.

This paper reviews and analyzes the provinces / municipalities (referred to as “region” in the later paper) in China based on the daily immigration population data and the daily number of newly diagnosed people. (Because Hong Kong, Macao, and Taiwan due to unavailability of immigration population data, and Hubei because of the origin of the virus and its traffic closure leads to little incoming immigration population from late January, these regions haven’t been analyzed. However, the immigration population moved from Hubei to other regions where are analyzed are considered.) The mathematical model established in this paper can be used to evaluate effectiveness of epidemic prevention and control of local governments in China. It can be used as a reference for other countries that are conducting epidemic prevention and control.

1. **Data used by the model**

This paper mainly uses two types of raw data. The first data is the daily number of newly diagnosed people in each region, which is referred to as “daily new diagnosed” or just “new” in the later paper (see Section 3.1 for details). The second data is the daily incoming immigration population in each region, which are distinguished from different source regions (see Section 3.2 for details). And the third type of data is constructed from the first two types of data, that is the risk of daily incoming immigration population in each region, which is referred to as “daily immigration risk” or just “risk” in the later paper (see Section 3.3 for details). This section describes these data.

* 1. **Daily new diagnosed in each region**

Since January 2020, various provinces / municipalities in China have successively announced data such as the number of new confirmed diagnoses of the epidemic each day. The “daily new diagnosed” data is crawled from Doctor Lilac Network (<https://ncov.dxy.cn/>).

* 1. **Daily incoming immigration population in each region**

The epidemic situation is exactly during the China Spring Festival transportation and its return journey. The population migration factor and its control must be the most important factor affecting the developing trend of the epidemic situation. Therefore, the daily incoming immigration population data, which are distinguished from different source regions, is crawled from Baidu Migration (<https://qianxi.baidu.com/>).

* 1. **Daily immigration risk in each region**

Due to the immigration population from different sources, the degree of risk is different. For example, people from Hubei and have a greater risk than other regions with the same immigration size. Therefore, process the incoming immigration population as follow, and get the “daily immigration risk” data.

riski = Σ (populationj \* newj)

riski is the “daily immigration risk” of region i. (i can be 1, 2, 3, …, 30. Because this paper analyzed 30 regions in China.)

populationj is the daily incoming immigration population from source region j. (j can be 1, 2, 3, …, 30, and j cannot be same as i.)

newj is the new diagnosed people count in source region j in recent 3 days.

The "recent 3 days" in newj is derived from the following: through testing, it is found that the fewer days used recently, the greater the correlation, and it is judged that the number of long-term confirmed diagnosis is difficult to reflect the current risk from source regions. But considering the number of new diagnosed people in single day fluctuates greatly, and the incubation period has a deviation of a few days. Therefore, "recent 3 days" is used here.

On the other hand, I also have considered the value in the brackets of the above formula, need to be divided by the total population of the source region before calculate cumulative, it means that when the value of newj in two region is same, the people from the region has smaller total population has greater probability as the virus transmitter moving into the destination region. However, after the correlation analysis of the subsequent models, it was found that the correlation coefficient did not improve, but the value of offset and window (see Section 4 for details) are increased to fit the similar correlation coefficient. It is judged that it is easier to be infected with the virus for migrants than non-migrants. So the above formula was used in the end.

1. **Model and data analysis**

The model in this paper uses the combination of the “offset” and time “window” parameters for the best fit of “daily new diagnosed” and “daily immigration risk”. And these two parameters serve as an evaluation for the effectiveness of external and internal epidemic prevention and control. This section performs the corresponding theoretical analysis and model parameter description.

* 1. **Calculation of correlation coefficient**

The model uses the data described in sections 3.1 and 3.3. Correlation analysis of these two sets of daily data is performed for each region in each day from January 17, 2020.

The analysis found that if the correlation coefficient is directly calculated for “daily new diagnosed” and “daily immigration risk”, its value may not be high enough, such as the correlation coefficient of Jiangsu Province is 0.675 (see Figure 1). There are even almost completely uncorrelated, such as the Heilongjiang Province. The coefficient is only 0.028 (see Figure 2). (Because the value of “daily immigration risk” is a very large value, drawing it with “daily new diagnosed” will make “daily new diagnosed” polyline draw as a very low horizontal line. Therefore, when drawing, the risk coefficient was adjusted, and the maximum “risk” value is set to the maximum “new” value, and the remaining “risk” values are reduced in proportion.)

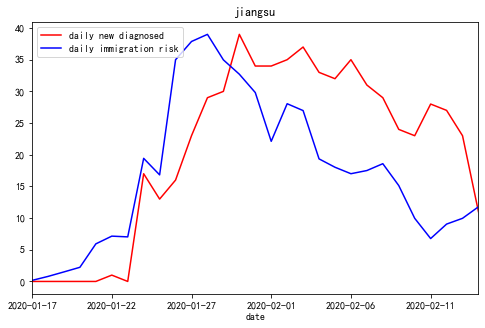


Figure 1. The correlation coefficient of Jiangsu Province is 0.675 from 2020-01-17 to 2020-02-14, when offset=0,window=1

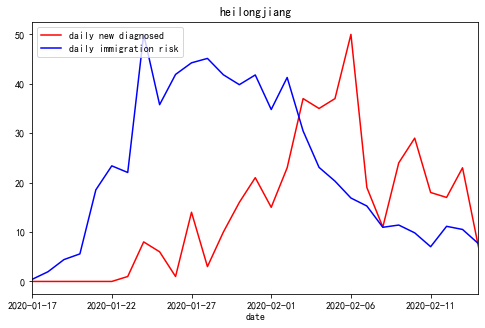


Figure 2. The correlation coefficient of Heilongjiang Province is 0.028 from 2020-01-17 to 2020-02-14, when offset=0,window=1

* 1. **Two model parameters: offset, window**

From the aspect of epidemic spread and prevention, there are two considerations as follows, and the two parameters in the models are introduced accordingly.

* + 1. **Control for incoming immigration population and “offset” parameter**

On the one hand, for the incoming immigration population, the preventive measures will be different at different times or in different regions. For example, in some areas, the temperature of immigrants, cough and other medical examinations have been strictly investigated relatively early. Such strict external people flow control will cause infected immigrants are immediately diagnosed, so the number of “daily new diagnosed” will increase significantly on the day when the “risk” factor is high or the day after. On the contrary, these externally infected people will be finally diagnosed relatively later. Therefore, the concept and the parameter of “offset” is introduced: the “risk” is shifted by a certain number of days, for example 3 days, which means that the “new” of each day is compared with the “risk” of 3 days ago. The calculation in Section 4.1 is calculated under offset = 0.

* + 1. **Control for inner people and “window” parameter**

On the other hand, the control of internal people and their own awareness of prevention will also affect the spread of the epidemic. For example, infected people who have moved inside the region (have been sick or incubated) and did not seek medical treatment or home isolation, and poor awareness such as going out without a mask will delay the external “risk” until the next few days before the “new” diagnosed people appear. In other words, the regions have the phenomenon of person-to-person transmission. On the contrary, only externally infected people are infected themselves, there will be only “offset” alone. Therefore, the concept of time window (“window”) is introduced into the model. For example, the “window” is 10, which means that the total “risk” for 10 consecutive days will affect the number of confirmed patients on the 10th day. From reference [1], we can know the incubation period with 95% confidence interval is 4.1-7.0 days, so if the infected person who moved in 10 days ago can still affect the inside, it should be much serious from person to person. The calculation in section 4.1 is calculated under window = 1.

* 1. **Use “offset” and “window” to process “risk” and calculate correlation coefficients**

Based on the analysis and the concepts introduced in Section 4.2, we set offset = 0, window = 9 and offset = 6, window = 5 in Jiangsu and Heilongjiang in Section 4.1, and calculated correlation coefficients of 0.952 and 0.865, as shown in Figures 3 and 4. The correlation coefficient is significantly improved, and the two line charts are significantly more fitted.

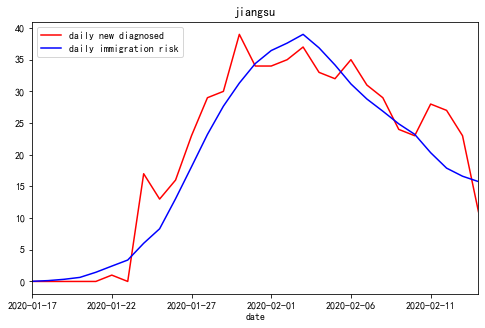


Figure 3. The correlation coefficient of Jiangsu Province is 0.952 from 2020-01-17 to 2020-02-14, when offset=0,window=9

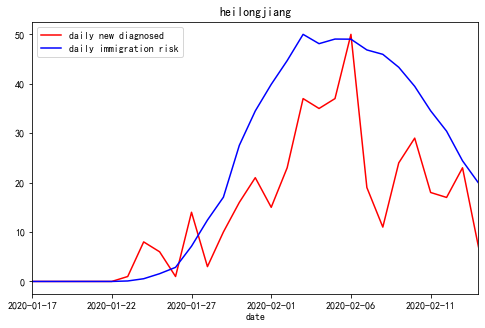
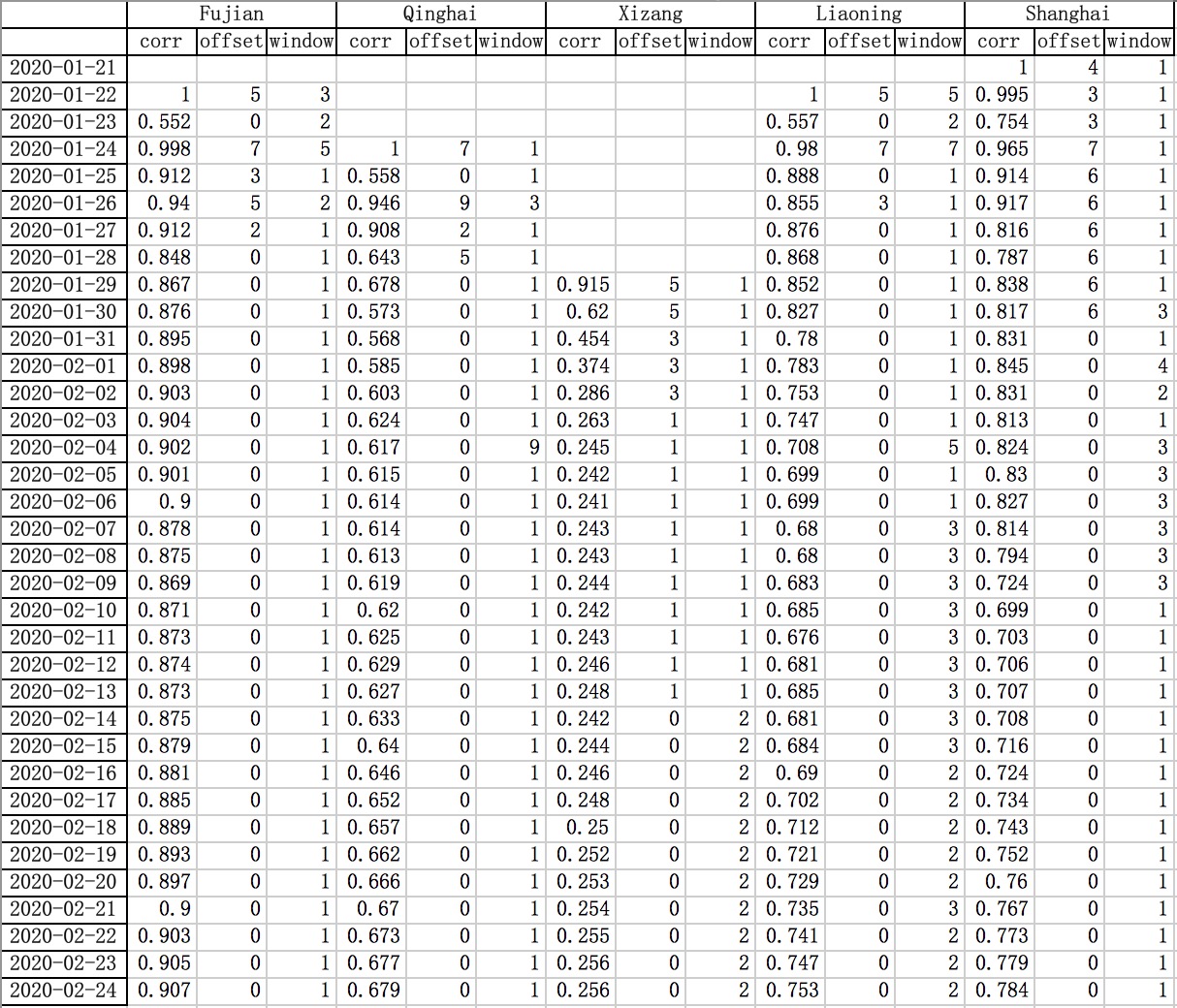
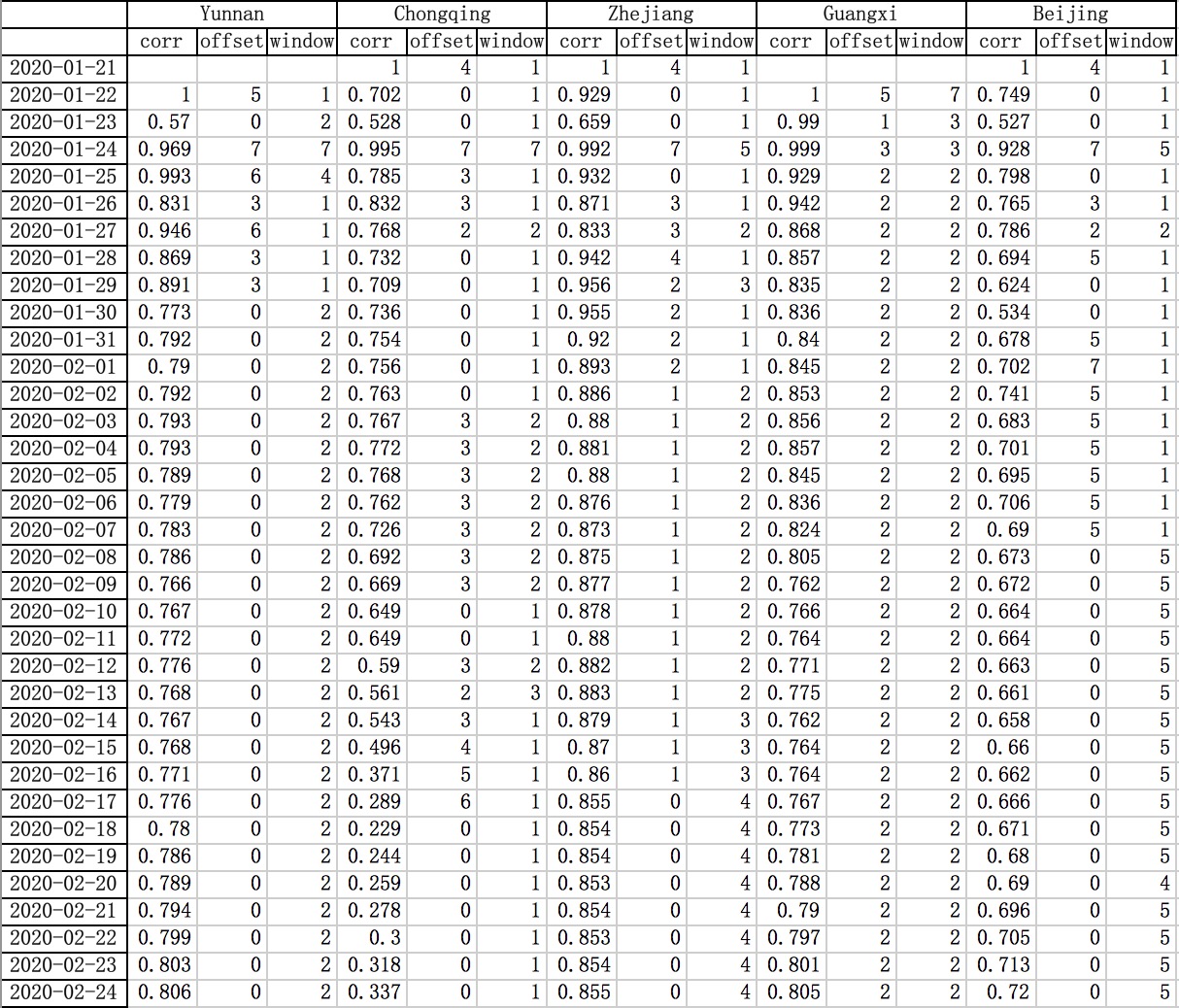


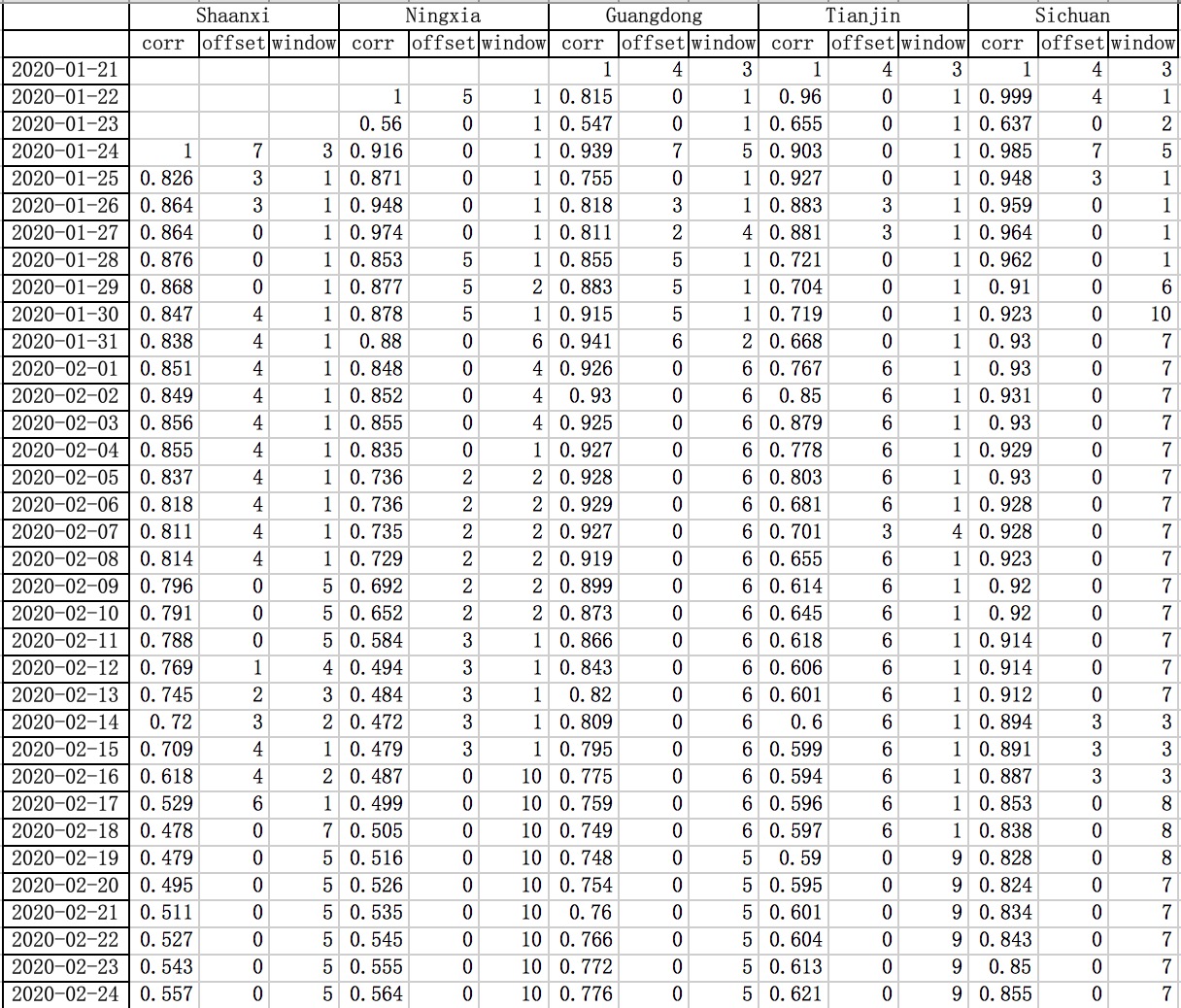
Figure 4. The correlation coefficient of Heilongjiang Province is 0.862 from 2020-01-17 to 2020-02-14, when offset=6,window=5

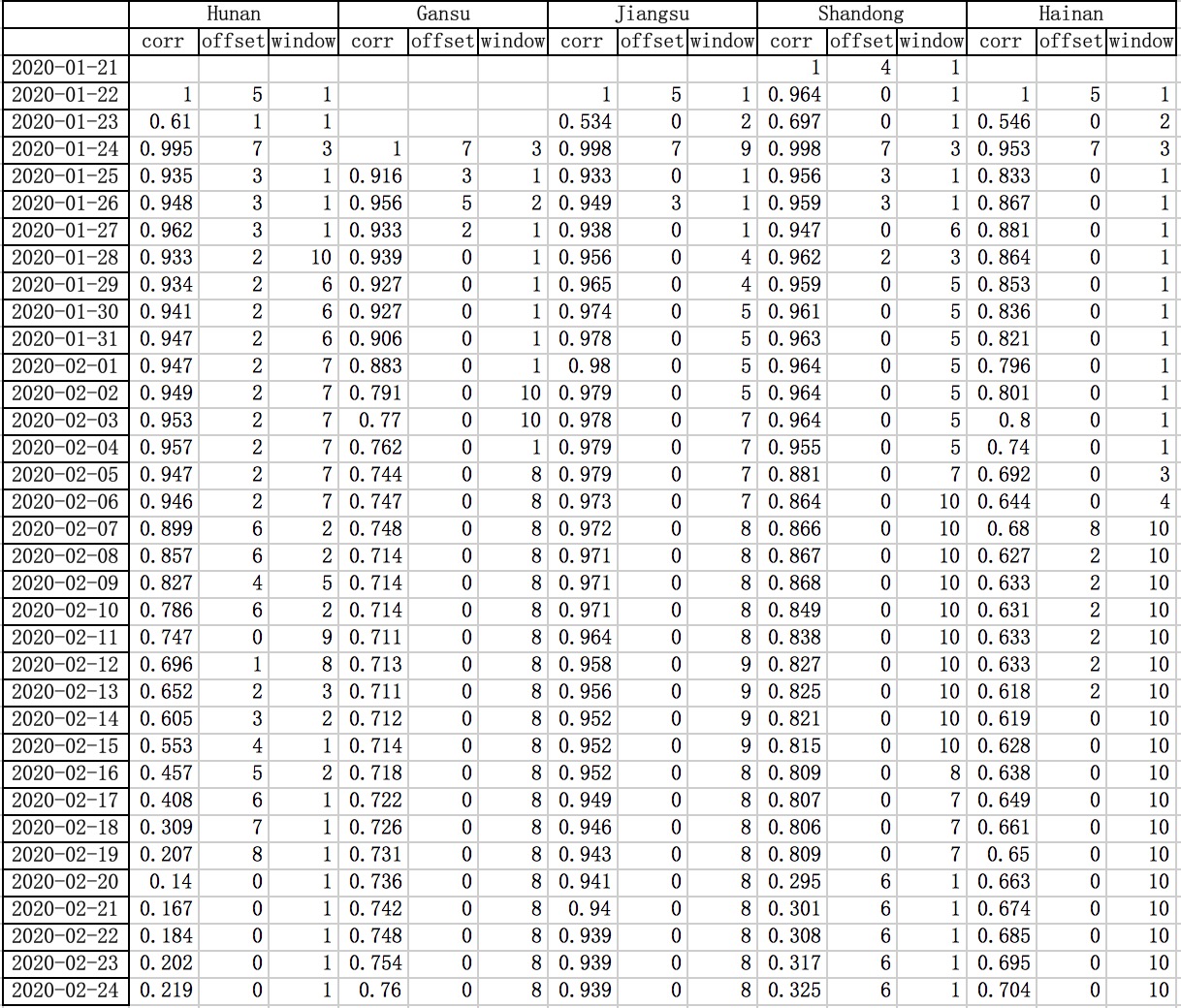
* 1. **Evaluation of the effectiveness of daily epidemic prevention and control in various regions of China**

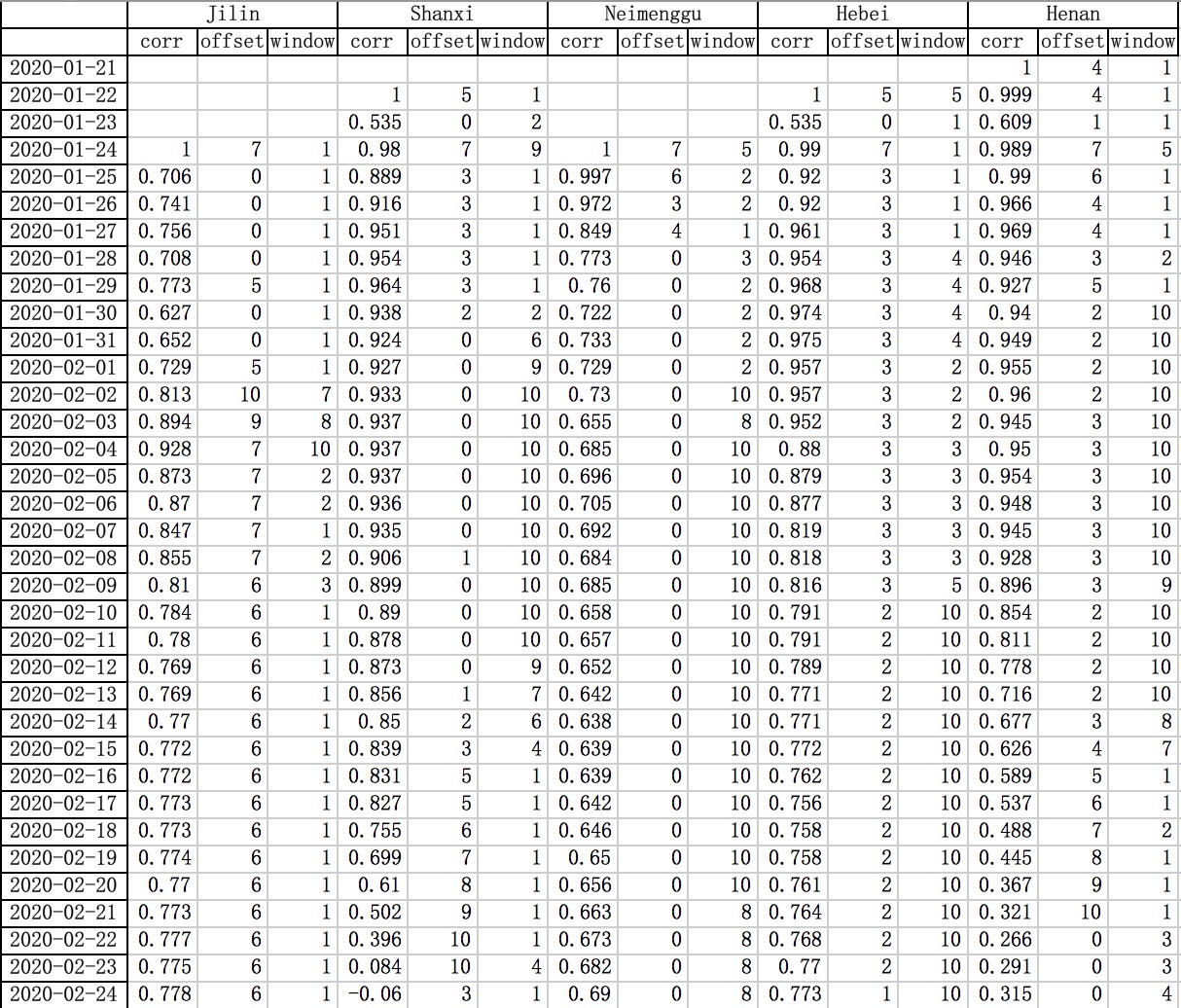
For all regions, starting from January 17, 2020, the “offset” is set from 0 to 10, and the “window” is set from 1 to 10. Each region has a total of 110 different “offsets” and “windows” combination. Calculate the 110 correlation coefficients with “daily new diagnosed” and “daily risk”, and finally take a set of “offset” and “window” corresponding to the maximum correlation coefficient. The result is shown in Table 1. This table can be used as an assessment of the effectiveness of prevention and control of the daily epidemic in each region. The regions are sorted by the average values of “offset” and “window” in ascending order. The blank spaces indicate that there were no diagnosed people in the regions on that day. The calculation results are available from the first day when the number of newly diagnosed people is not zero.











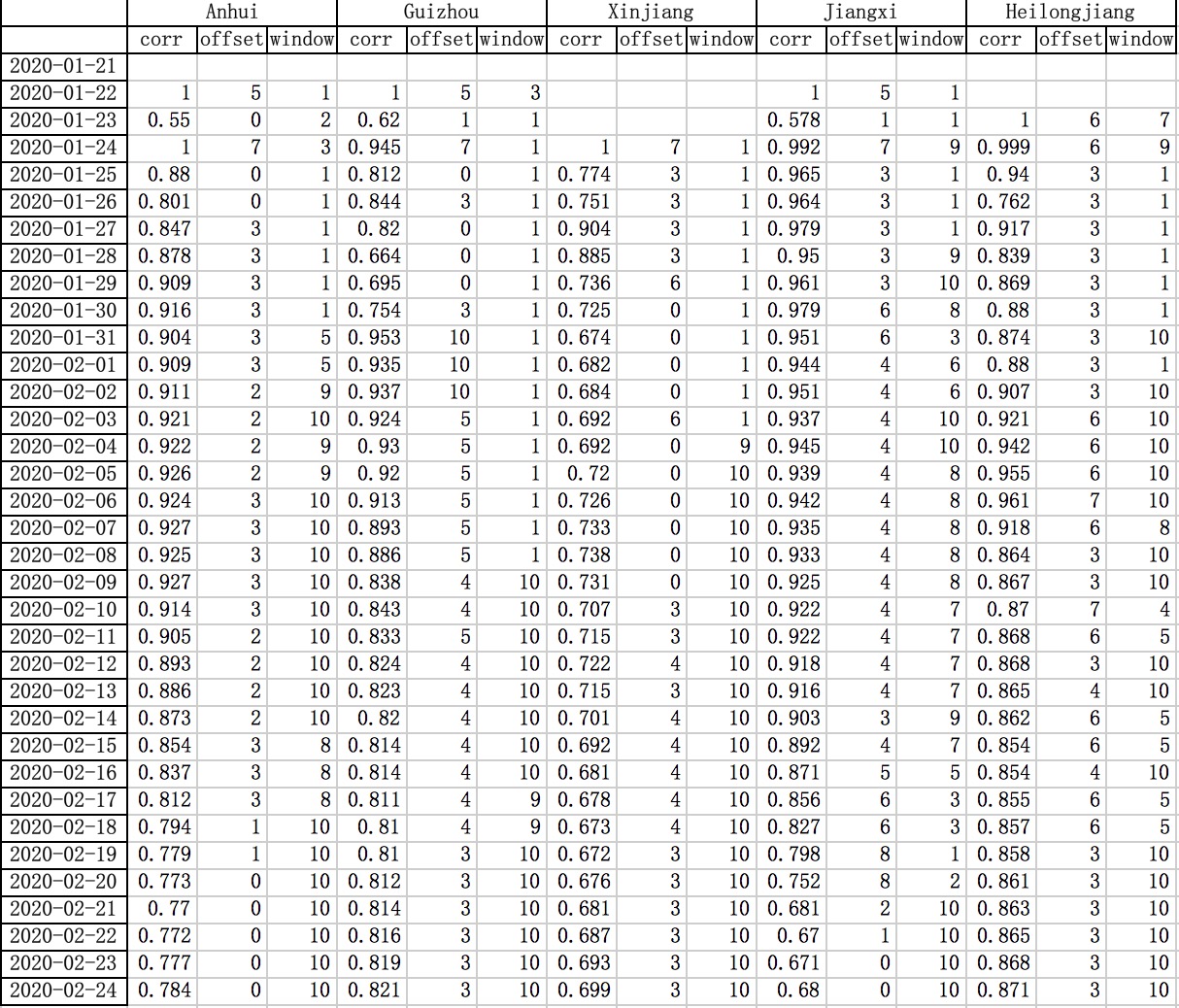


Table 1. Model evaluation on the effect of prevention and control in each region of China

* 1. **Analysis of model evaluation results**

From the analysis in Section 4.2, it can be known that a large “offset” value indicates that the incoming immigration population check and control (external prevention and control) is relatively loose, otherwise it is strict; a larger window value indicates that the internal prevention and security awareness (internal prevention and control) is relatively weak, and vice versa.

For example, we can draw conclusions from the values of “offset” and “window” in various regions:

In Fujian and some other regions, internal and external prevention and control are very effective. The values of offset and window are both small;

In Jiangsu and some other regions, although the flow of people moving in from outside has been well controlled, internal prevention and control or the safety awareness of internal personnel need to be strengthened. The values of offset are small, but the values of window are somewhat large;

In Jilin and some other regions, although internal prevention and control or internal personnel's safety awareness is strong, the external flow of people needs to be strengthened. The values of offset are somewhat large, but the values of window value are small;

In Heilongjiang and some other regions, internal and external prevention and control needs to be strengthened. The values of offset and window are both large.

In addition, we can see more information from the values of “corr” in the table. It is the correlation coefficient corresponding to “offset” and “window”, and is the highest “corr” value of the 110 “offset” and “window” sets. As the number of days of epidemic situation continues to increase, the overall trend of “corr” values in various regions gradually becomes smaller. This means that the confidence of the corresponding “offset” and “window” values is reduced. The analysis found that there are some details as follow:

First, the calculation of the correlation coefficients for each day starts from the data on January 17th. However, in early February, the values of “daily new diagnosed” (except Hubei) was the largest. From around February 20th, the values of “daily new diagnosed” (except Hubei) was turning to 0 or close to 0;

Second, from around February 10th, although Chinese Spring Festival holiday is over, the number of migrants between regions is not large, much fewer than last year's Spring Festival return trip;

Third, in this outbreak, there was only one person who was diagnosed with the infection in Xizang (Tibet). And looking at the migration data, it is found that the number of people entering Xizang was also very small, so the correlation coefficient has been very low.

Therefore, from the above details, the reason can be judged: when the values of “daily immigration risk” and “daily new diagnosed” are small, there is not enough sample data input into the model, the correlation coefficient value becomes small. Because the “corr” value is proportional to the confidence of fitting the “risk” and “new” values, the smaller the number of samples, the greater the chance, and the smaller the statistical confidence. Therefore, when the values of “risk” and “new” are small, the accuracy of the model evaluation results will decrease. Therefore, the date with the highest model confidence is before February 15th, in the rough estimation.

1. **Relevant news and data confirm the model**

The following news and data confirm the accuracy of the model.

First, according to the diagnosis data released by Doctor Lilac Network in China, the numbers of cumulative diagnosed people are divided into incoming immigrants and local people in only 3 regions: Shanghai, Beijing, and Tianjin. As of the afternoon of February 25, the incoming/local cumulative number of diagnosed people in these three regions is 111/ 225, 25/375, 6/129, that is, Shanghai's control is best (local people infection rate is lowest), Beijing is second, and Tianjin is relatively poor. From the values of offset and window of these three regions, it is also the lowest in Shanghai, followed by Beijing, and Tianjin is highest.

Second, on February 8th, reference [2] pointed out that Heilongjiang did not pay enough attention to the epidemic and had poor awareness of prevention. Judging from the offset and window values of Heilongjiang, this is indeed the case.

Third, the author of reference [3] collected R0 data from Shanghai, Zhejiang, Jiangsu, Anhui, Henan, and Sichuan (excluding Hubei, Wuhan, and China, which are not counted in this paper). The R0 values are ranked in ascending order as follows: Shanghai , Zhejiang, Henan, Sichuan, Jiangsu, and Anhui, which are almost the same as the rankings in Table 1, with only difference in Henan. The control effects of the 6 regions in Table 1 from good to poor are: Shanghai, Zhejiang, Sichuan, Jiangsu, Henan, Anhui. It is speculated that there may be discrepancies in clinical R0 indicators in individual regions.

Last, in late January, there were a large number of infected businessmen who were returning to Wenzhou, Zhejiang from Wuhan, Hubei. In reference [4], on February 1th, the government of Wenzhou, Zhejiang issued 25 strict prevention and control measures in a timely manner, and finally the Wenzhou epidemic situation was well controlled, so the prevention and control effect in Zhejiang Province, where Wenzhou City is located, is very good. Although the cumulative number of diagnoses is large in Zhejiang, the values of “offset” and “window” are relatively low.

1. **Summary**

This paper analyzes the “daily new diagnosed” and “daily incoming immigration population”, and puts forward the concept and calculation method of the “daily immigration risk” factor, and uses the concept of “offset” and “window” to fit the data and use it to evaluate the effect of external and internal epidemic prevention and control.

At present, although the epidemic prevention and control in China has achieved great results, it continues. New outbreaks are gradually emerging around the world. The model in this article should be applicable to the analysis of epidemic prevention and control effects in various parts of the world. I hope this article can contribute to the prevention and control of the world epidemic. It ’s like a saying in Chinese, we live in a global village, and we are a family. We Chinese people hope that the epidemic will be brought under control worldwide as soon as possible and finally ended.

1. **References**

[1] Qun Li, M.Med., Xuhua Guan, Ph.D., Peng Wu, Ph.D., Xiaoye Wang, M.P.H., Lei Zhou, M.Med., Yeqing Tong, Ph.D., Ruiqi Ren, M.Med., Kathy S.M. Leung, Ph.D., Eric H.Y. Lau, Ph.D., Jessica Y. Wong, Ph.D., Xuesen Xing, Ph.D., Nijuan Xiang, M.Med., et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. The New England Journal of Medicine, January 29, 2020 DOI: 10.1056/NEJMoa2001316. (2020-01-29). <https://www.nejm.org/doi/full/10.1056/NEJMoa2001316?query=featured_home>.

[2] Why Heilongjiang has the highest number of confirmed diagnoses in northeast of China. Doctor Lilac Network. (2020-02-08). <https://mama.dxy.com/outbreak/daily-of-nationwide-new?index=20200208&locationIds=999&share=true&entry=Previous_Review>

[3] Qingzhu Dai. When will the epidemic end? The reliable answer is here!. alibaba staff blockchain alliance.(2020-02-11).<https://mp.weixin.qq.com/s/6I1onP3jt3G76IYjy9YWlw>

[4] Wenzhou comprehensively implements "25" emergency measures for epidemic prevention and control. Wenzhou government. (2020-02-01). <http://www.wenzhou.gov.cn/art/2020/2/1/art_1217828_41865258.html>