International Journal of Statistical Distributions and Applications
2023; 9(1): 41-48
http://www.sciencepublishinggroup.com/j/ijsda
doi: 10.11648/j.ijsd.20230901.15
ISSN: 2472-3487 (Print); ISSN: 2472-3509 (Online)

SciencePG
Science Publishing Group

Bayesian Analysis on the Spatial Difference of Input Risk of Overseas Cases of COVID-19 in China

Bo Yang[1], Yunyuan Yang[2], Wei Zheng[1], Yanmei Li[1], Xinping Yang[1, *]

[1]School of Mathematics and Computer Science, Chuxiong Normal University, Chuxiong, China

[2]School of Environment and Chemistry, Chuxiong Normal University, Chuxiong, China

Email address:

Yangxp@cxtc.edu.cn (Xinping Yang)

[*]Corresponding author

To cite this article:

Bo Yang, Yunyuan Yang, Wei Zheng, Yanmei Li, Xinping Yang. Bayesian Analysis on the Spatial Difference of Input Risk of Overseas Cases of COVID-19 in China. *International Journal of Statistical Distributions and Applications*. Vol. 9, No. 1, 2023, pp. 41-48.
doi: 10.11648/j.ijsd.20230901.15

Received: February 24, 2023; **Accepted**: March 20, 2023s; **Published**: March 31, 2023

Abstract: To analyze the spatial difference of COVID-19 import risk is helpful for scientific prevention and control. On the basis of clustering 25 provinces and cities with epidemic input in study time, a multinomial distribution model was established under the Bayesian framework. All parameters Bayesian estimation was obtained by MCMC method. 25 provinces and cities with overseas input were divided into 9 categories from March 3 to April 23, 2020. 468 overseas input risk values are regarded as parameters, and the maximum MC-error estimated by Bayesian is only 0.677% of the standard deviation. During the study period, 25 provinces and cities have input risk. The highest risk areas of overseas import are 12 provinces and cities in the first category represented by Beijing, Shanghai and Guangdong Province, including 10 provinces and cities along the coast / border. The lowest risk areas are the eighth category (Henan Province) and the ninth category (Anhui Province); the fourth category (Heilongjiang Province and Shanxi Province) risk is higher than the first category in 7 days and it has the largest input vary fluctuation. Taking 2020-3-22, 4-7 and 4-18 as time nodes, the overseas input risk is divided into four stages. In the first stages, the highest risk of overseas import is the first category (59.613%); in the second and third stages are the first category (decline from 60.505% to 37.056%), the fourth category (increase from 16.071% to 33.852%); in the fourth stage, the first category (42.622%), the third category (Shaanxi Province and Jilin Province, 17.556%) and the fourth category (10.056%).

Keywords: COVID-19, Overseas Input Risk, Multiple Distribution, MCMC Method, Bayesian Estimation

1. Introduction

As of 2020-3-28, there were 11 days in Wuhan, the worst affected area of the epidemic in China, with 0 new cases. On the same day, the Ministry of Foreign Affairs and the State Administration of Immigration issued a notice suspending the entry of foreigners holding valid Chinese visas and residence permits from 0:00 on March 28. 2020-4-8, Wuhan lifted the blockade, but the number of imported cases abroad was 1103. By 2020-4-23, 1618 cases had been recorded for overseas imports. China's epidemic prevention work has shifted to the importation of overseas cases (including overseas importation of associated cases) and asymptomatic infection prevention and control. Therefore, the spatial difference analysis of the risk of overseas imported cases can serve the decision-making of epidemic prevention and control in various provinces (cities).

The existing quantitative study of COVID-19 mainly focuses on the description statistics of epidemic trends [1, 2] and trend estimation. For examples, estimates of the size of the COVID-19 outbreak were made using SEIR or modified SEIR models [3-5], self-regression moving average models (ARIMAs) [6], random transmission models [7], etc; COVID-19 regeneration estimation was carried out using the Marcof Monte Carlo Method (MCMC) [8-9] and so on. Bayesian method has some advantages in data analysis in the field of medicine, the uncertainty parameter value can be quantified under the premise of known data, and the quantitative accuracy of parameters can be improved by a priori information and data information [10]. For example, Han Ke et al. [11] used the Poisson distribution model under bayesian framework to estimate the number of COVID-19 regenerations in first-tier cities. On the study of spatial differences in the COVID-19 outbreak, the Johns Hopkins Center for Systems Science and Engineering produced a map of the global outbreak [12], with a maximum of more than 2 billion daily visits. Guan Weijie etc [13] and Qi Cuifang etc [14] carried on the study of nationwide clinic characteristic of epidemic situation and COVID-19 inter-provincial communication and influencing factor analysis respectively, and both of them used GIS mapping to reflect the distribution of confirmed cases in each province and city. Yi Dali [15] and others carried out cluster analysis with the epidemic data in 34 provinces and cities from 2020-1-19 to 2-16, a total of 6 categories, among them, it was the high risk areas of Hubei Province and Henan Province that needed to be strictly controlled. According to Tencent location data and Baidu migration data, Liu Zhang [16] et al. have completed the spatial distribution estimate of people who moved out of Wuhan during the COVID-19 outbreak. In the scale of Wuhan City 1140 traffic analysis area, Feng Mingxiang [17] and others carried out COVID-19 space-time diffusion estimate combined with mobile phone user space interaction data. GIS, multi-source data and big data platform are effective methods for the study of COVID-19 outbreak simulation and spatial distribution differences. However, the analysis and processing of zero expansion data, missing data and short-term data by these methods often result in a great deviation from the actual situation. At present, foreign imported case data in China have the characteristics of zero expansion, geographical absence, data size differences, the result may deviate from the actual situation of the current domestic and foreign input if using the traditional method for analysis and processing. Based on the number of overseas imported cases from 2020-3-3 to 4-23 and the clustering results, this paper constructs a multinomial distribution model of the probability of input risk in each province and city under the framework of Bayesian, solves the model by MCMC method, and analyzes the spatial differences in the input risk of cases of COVID-19 outside China, The research results are expected to serve for epidemic prevention and control abroad.

2. Data and Methods

Collect daily new case data from 25 provinces (cities) in China involved in overseas imported cases from 2020-3-3 to 4-23 as a sample (data from the Bulletin of the National Health And Health Commission and the Provincial and Municipal Health and Construction Commission). The zero expansion characteristics of this data are obvious, the daily data of provinces and cities are different, and the regional differences are obvious. For example, among 25 provinces and cities, Heilongjiang imported 86 cases in 2020-4-7, and it was the largest number of imported cases day by day, however there were 0 imported cases in 18 provinces in the same day. From the frequency (proportion) point of view, the frequency of foreign import cases was 0, but this could not be explained that there were no risk of overseas import in 4-7 in these 18 provinces and cities. If we choose the methods used in the literatures to deal with these data, it inevitably results in most time points not in line with the actual situation.

By using two clustering methods to cluster data from 25 provinces and cities with overseas outbreak input in China, the results can be obtained relatively close. Based on the cluster results, the model of the probability of input risk abroad under Bayesian is established, the appropriate Dirichlet distribution is used as its priori distribution, the model parameters (input risk) are solved by MCMC method, and the GIS mapping is used to reveal the spatial difference of the input risk of COVID-19 cases abroad. The software used are R language and OpenBUGS, and the map mapping software are Adobe Illustrator and Photoshop.

3. Model of Import Risk for Overseas Cases

3.1. Clustering of Imported Cases from Abroad

Programs are written in R language to cluster sample data (2020-3-3 to 4-23) from 25 provinces and municipalities. The name of the province and city is considered an indicator, the similarity coefficient between the two indicators $X_i, X_j, (i \neq j)$ is defined as $c_{ij} = \frac{X_i^T X_j}{\|X_i\|_2^{1/2}\|X_j\|_2^{1/2}}$, among $\|\cdot\|_2^{1/2}$ represents the square root of two norm of the vector [18], and the distance between the two variables is $d_{ij} = 1 - c_{ij}$.

Usually the results of different clustering methods are different, and the determination of the number of classifications is not yet fully resolved. In the actual study, according to the purpose of the study, we generally choose a variety of methods of clustering, through comparative analysis, to determine the final clustering method and classification number to get better results [18].

Through the comparative analysis of various clustering results, it is found that the clustering results of the class average method (Average) and the similar method (Mcquitty) are close, and the average method of the class is not concentrated or expanded, which is the clustering method recommended by many literatures [19]. By combining the advantages of the two clusters, the following clustering results are obtained:

Category I (G_1): Zhejiang, Beijing, Shanghai, Guangdong, Sichuan, Shandong, Yunnan, Guangxi, Fujian, Tianjin, Liaoning, Jiangsu, a total of 12 provinces and cities. The corresponding sample data from 2020-3-3 began to form a data matrix of the lower trapezoidal structure, the largest number of cases, up to 917 cases.

Category II (G_2): Jiangxi, Chongqing, Guizhou, 3 provinces and cities. Input cases were concentrated in 3-21 to 3-28, and the number was small, only 6 cases.

Category III (G_3): Shaanxi, Jilin, 2 provinces. Input cases were concentrated in 3-21 to 4-23, with continuous importation of overseas cases, a total of 49 cases.

Category IV(G_4): Heilongjiang, Shanxi, 2 provinces. Input cases were concentrated in 3-18 to 4-23, with continuous importation of overseas cases, a total of 445 cases.

Category V (G_5): Hebei, Hunan, 2 provinces. Input cases were concentrated in 3-21 to 4-15 cases, a total of 11 cases.

Category VI (G_6): Inner Mongolia. Input cases were concentrated in 3-24 to 4-15 cases, a total of 118 cases.

Category VII (G_7): Gansu. Input cases were concentrated in 3-5 to 4-5 cases, a total of 47 cases.

Category VIII (G_8): Henan. Input cases were concentrated in 3-11 to 3-25 cases, a total of 3 cases.

Category IX (G_9): Anhui. Only one input case in 4-8.

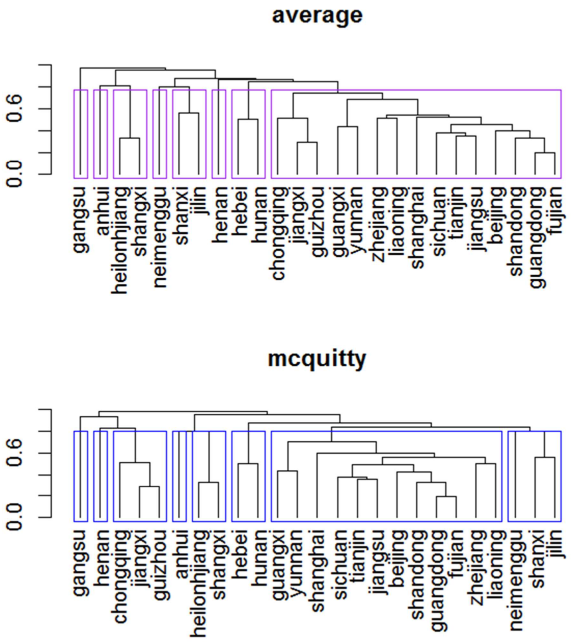


Figure 1. *Cluster tree of daily number of imported cases of 25 provinces or cities from March 3 to April 8, 2020.*

Note: Use Chinese Pinyin as the variable name of 25 provinces or cities, in order to avoid the same name in Pinyin, use shangxi to represent Shanxi

3.2. Bayesian Model

3.2.1. Construction of Model

Any confirmed case of new import outside the province *i* of *t*-day is indicated by b_{ti}, the oe_t indicates the number of new cases imported from outside on the *t*-day, and the number of cases entered from the province of category j on t-day is indicated by g_{tj}, then $oe_t = \sum_{i=1}^{25} b_{ti} = \sum_{j=1}^{9} g_{tj}$. The risk of foreign input for category *j* in *t*-day is defined as: $P\left(b_{ti} \in G_j \middle| i\ province \in G_j\right) = p_{tj}, \sum_{j=1}^{9} p_{tj} = 1$.

This probability expresses the risk of foreign input for category *j* in *t*-day. The higher the value, the greater the risk of the overseas COVID-19 input of category *j* on *t*-day, the greater the pressure of prevention and control from overseas input. The number oe_t of new overseas imported cases in the nine categories on day *t*-day is considered as oe_t independent tests, and each confirmed case imported from abroad is considered to be a test. The test result can only belong to one of the nine categories, the risk probability of the case belonging to category j is p_{tj} and it is an evaluatable parameter. Vectors consisting of the parameters to be evaluated and data of new overseas imported cases in the nine categories on *t*-day are recorded as θ_t and $data_t$, namely:

$$\theta_t = (p_{t1}, p_{t2}, p_{t3}, p_{t4}, p_{t5}, p_{t6}, p_{t7}, p_{t8}, p_{t9})^T, data_t = (g_{t1}, g_{t2}, g_{t3}, g_{t4}, g_{t5}, g_{t6}, g_{t7}, g_{t8}, g_{t9})^T.$$

By the definition of multinomail distribution, we get $data_t \sim M(oe_t, \theta_t)$ and the likelihood function is:

$$p(\theta_t|data_t) = \Gamma(oe_t + 1)\prod_{j=1}^{9} \frac{p_{tj}^{g_{tj}}}{\Gamma(g_{tj}+1)}, \quad (1)$$

where $\Gamma(\gamma) = \int_0^\infty x^{\gamma-1}e^{-x}dx$. According to proposals such as NguyenX (2016) [20], the Dirichlet distribution is used as a priori distribution, and the number of provinces and cities contained in each category in the clustering results is used a priori information to select prior parameters, namely, $\alpha = (12,3,2,2,2,1,1,1,1)^T$, The priori density is $p(\theta_t) \propto \prod_{j=1}^{9} \theta_{tj}^{\alpha_j - 1}$, and we obtain the post-test exact distribution of parameter θ_t:

$$\theta_t \sim Dirichlet(\alpha_1 + g_{t1}, \alpha_2 + g_{t2}, \alpha_3 + g_{t3}, \ldots, \alpha_9 + g_{t9})$$

and the corresponding post-test density function is

$$p(\theta_t|data_t, \alpha) = \Gamma(oe_t + 25)\prod_{j=1}^{9} \frac{p_{tj}^{g_{tj}+\alpha_j}}{\Gamma(g_{tj}+\alpha_j)}. \quad (2)$$

By (2) we can get the full condition distribution of nine parameters, for example, in the case of θ_{t1}, the corresponding full-condition distribution is

$$p\left(\theta_{t1}|\theta_{t(-1)}, data, \alpha\right) \propto p_{t1}^{12+g_{t1}} = Beta(13 + g_{t1}, 1). \quad (3)$$

Formula (3) is a standard distribution, the other eight full-condition distributions are also beta distribution. We can use Gibbs sampling to obtain all the parameters (a total of 52 days, 9 categories per day, 52 × 9 = 468 parameters) of the post-test sample, and complete Bayesian inference.

3.2.2. Solution of the Model

The solving of the model is completed in the OpenBUGS software environment. After seeding random numbers, the software automatically generates the initial value of 468 parameters. In order to reduce the self-correlation among the post-test samples and to ensure the convergence of the MC chain, the sampling step of the sampling interval is 100. The orderly loosening algorithm [21] is used to eliminate random walking in the MCMC. After 10^5 iterations, posterrior samples (468 MC chains) of 468 parameters to be evaluated was obtained. After each chain throws away the first 4999 samples of the burnin period, the parameters θ are inferred by the MC method using the remaining 29619 samples. The corresponding parameter estimates are shown in Table 1 (9 categories, 52 values per category). Table 1 includes mean estimates for each parameter (Mean), median estimates (Median), 95% confidence interval CI: (2.5% ql, 97.5%qu), standard deviation (SD), and MC error

(MCerror). See table 2 for the MC errors of category I to IX overseas input with greatest risk and the time point of occurrence. Each has a maximum MC error of less than 5.56×10^{-4}. The maximum of (*max.MCerror*/*sd*)*100%is 0.677% (far smaller than 10%), which means that the model (2) has high precision [22].

Table 1. *Summary of Bayesian estimation and some statistics of 468 parameters.*

node	Mean	Sd	MCerror	2.50%ql	Median	97.50%qu
p_{11}	0.6066	0.08335	0.000494	0.4399	0.6088	0.7621
p_{21}	0.5179	0.09425	0.000518	0.3310	0.5190	0.6995
…	…	…	…	…	…	…
$p_{52.1}$	0.4812	0.09475	0.000556	0.2978	0.4807	0.6674
node	Mean	Sd	MCerror	2.50%ql	Median	97.50%qu
p_{13}	0.06057	0.04082	0.000230	0.007569	0.05202	0.1612
p_{23}	0.07441	0.04936	0.000286	0.009637	0.06430	0.1960
…	…	…	…	…	…	…
$p_{52.3}$	0.1112	0.06001	0.000364	0.02456	0.1013	0.2547
node	Mean	Sd	MCerror	2.50%ql	Median	97.50%qu
p_{15}	0.06037	0.04082	0.000224	0.007498	0.05176	0.1623
p_{25}	0.07411	0.04924	0.000298	0.009677	0.06386	0.1953
…	…	…	…	…	…	…
$p_{52.5}$	0.07440	0.04953	0.000281	0.009567	0.06392	0.1978

node	Mean	Sd	MCerror	2.50%ql	Median	97.50%qu
p_{12}	0.09115	0.04916	0.000299	0.01986	0.08314	0.20840
p_{22}	0.1109	0.05922	0.000327	0.02440	0.10120	0.25110
…	…	…	…	…	…	…
$p_{52.2}$	0.1115	0.05994	0.000363	0.02396	0.10190	0.25210
node	Mean	Sd	MCerror	2.50%ql	Median	97.50%qu
p_{14}	0.06101	0.04122	0.000238	0.007796	0.05221	0.1637
p_{24}	0.07364	0.04914	0.000258	0.009641	0.06348	0.1945
…	…	…	…	…	…	…
$p_{52.4}$	0.07385	0.04962	0.000321	0.009074	0.06352	0.1968
node	Mean	Sd	MCerror	2.50%ql	Median	97.50%qu
p_{19}	0.02995	0.02898	0.000167	0.000757	0.02119	0.1073
p_{29}	0.03736	0.03615	0.000187	0.000988	0.02658	0.1345
…	…	…	…	…	…	…
$p_{52.9}$	0.03702	0.03574	0.000215	0.000999	0.02613	0.1327

Note: Categories 6-8 are omitted from the table

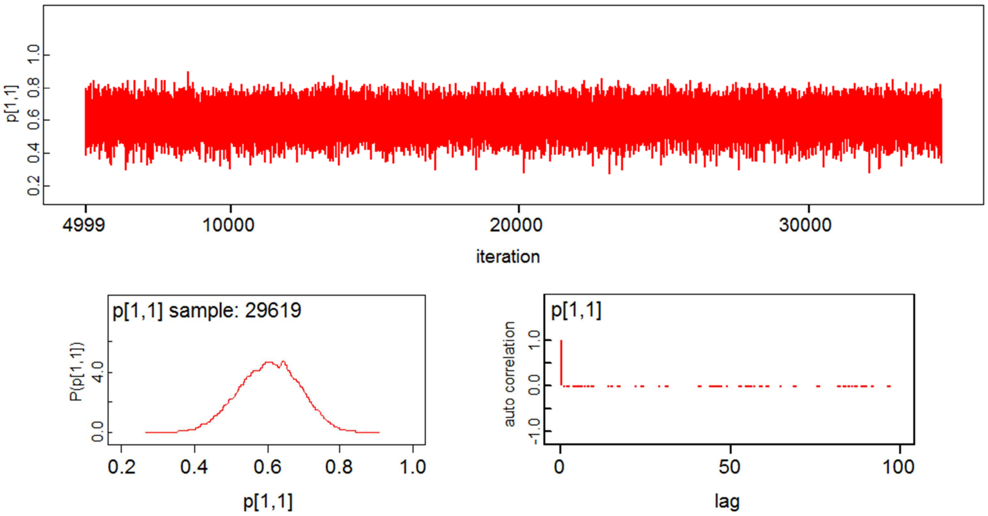


Figure 2. *Convergence diagnosis diagram. The upper is the history strace plot graph of posterior samples of p_{11}, the lower left is the curve graph of kernel density estimation of p_{11}, and the lower right is the autocorrelation graph of p_{11}.*

3.2.3. Diagnosis of Model

For the MC chain of 468 parameters in model (2), the History-strace-plot plot, the density estimation graph, and the Autocorrelation graph are drawn respectively (in the case of p_{11}, see Figure 1, the remaining 467 plots are omitted, the same below). The History-strace-plots of 468 parameters show that after discarding the first 4999 burnin values, the MC chain of 468 parameters converge and each limit distribution is their own posterior distribution. The autocorred graph of each MC chain shows that after the lag period ≥2, the autocorrelation coefficient is close to 0, and each MC chain can be regarded as a MC chain of the independent and identically distribution. Statistical inferences can be made using the corresponding MC chain as a posterior sample (see Figure 1). the category I $p_{t,1}$ presents symmetrical distribution characteristics (it is also proved by 52 box diagrams of the first category in Figure 2), and the probability of the other nine categories presents a more severe right-biased distribution, with mean estimates being more affected by extreme values. Their robust estimates (median value) and corresponding 95% confidence interval are taken as the risk probabilities of the nine categories for discussion.

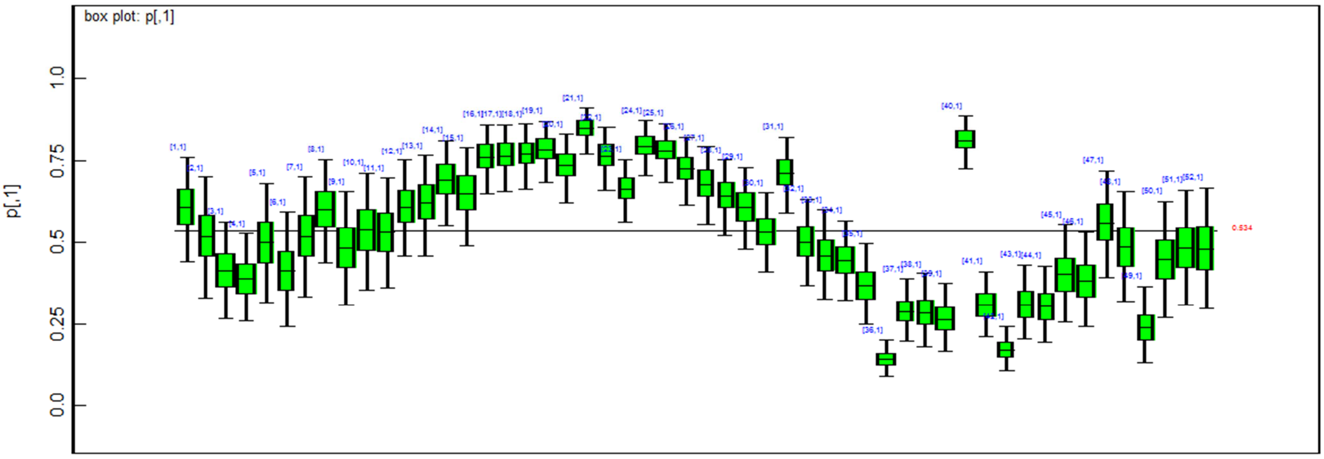


Figure 3. *The box plot of posterior sample of* p_{11} *to* $p_{52,1}$

Table 2. *Summary of maximum MC error and corresponding information of each MC chain of model (2).*

node	Max.MC error	Sd	Time point of occurrence	The frequency of the sample calculation	Prediction of the model (2)	(Max.Mc error/sd) 100%
p_{t1} of G_1	0.000556	0.09475	2020-4-23	0.5000	0.4807	0.587%
p_{t2} of G_2	0.000371	0.05730	2020-3-12	0.0000	0.09741	0.647%
p_{t3} of G_3	0.000408	0.07016	2020-4-20	0.0000	0.5004	0.581%
p_{t4} of G_4	0.000443	0.07363	2020-4-17	0.73333	0.3549	0.602%
p_{t5} of G_5	0.000301	0.04888	2020-3-9	0.0000	0.06333	0.616%
p_{t5} of G_5	0.000318	0.05884	2020-4-10	0.05263	0.4106	0.540%
p_{t5} of G_5	0.000432	0.06811	2020-3-6	0.6875	0.3656	0.6345%
p_{t5} of G_5	0.000276	0.04386	2020-3-11	0.0000	0.05557	0.629%
p_{t5} of G_5	0.000232	0.03429	2020-3-12	0.0000	0.02538	0.677%

Note: p_{ti} denotes the probability of G_i $(i = 1,2, \dots 9)$.

4. Interpretation of Result

The 52-day 2020-3-3 to 4-23 days are considered to be 52 time points, with 0 new cases of overseas input in nine categories with many time points. The frequency (proportion) of overseas input is calculated with sample data, and the input frequency of the corresponding time point is 0 or 1. However, this does not mean that the risk of overseas input in these points is 0 or that the risk of overseas input is 100%. According to probability theory, the estimation error is large at these time points corresponding to these extreme values [23]. The calculation frequency of each of the five time points (2020-3-9, 3-11, 3-12, and 4-20) in Table 2 is 0, which is also explained by the maximum MC error (the second column in Table 2). Using the model (2) calculated by the nine categories of overseas input risk (converted to percentages). The nine categories show the risk change characteristics of four stages. In order to clearly express the change characteristics of each stage, the four categories (G1, G2, G4, G6) with large overseas input risks and the remaining five categories are respectively plotted as point and line graphs. The statistical values of each category are calculated (Table 3). The results show that: (1) In the first stage (2020-3-3 (No. 1) to 3-21 (No. 21)), except for the rapid rise of category G1 and the rapid decline of category G7, other types of input risks decline slowly. (2) In the second stage (2020-3-22 (No. 22) to 4-7 (No. 39)), except for the rapid decline of category G1 and the rapid rise of category G4, other types of input risks rose slowly and gradually declined after seven days.(3) In the third stage (2020-4-8 (No. 40) to 4-18 (No. 49)), the G4 category of overseas input risks declined rapidly, while other types of input risks showed an obvious upward trend. (4) In the fourth stage (after 4-18 (No. 50)), the input risks of each category fluctuate steadily (Figure 3). The overseas input risk probability of nine categories is between 0 and 1, which indicates that the overseas input risk probability obtained by model (2) can truly reflect the spatial distribution of overseas input cases, which is more practical than the frequency to describe.

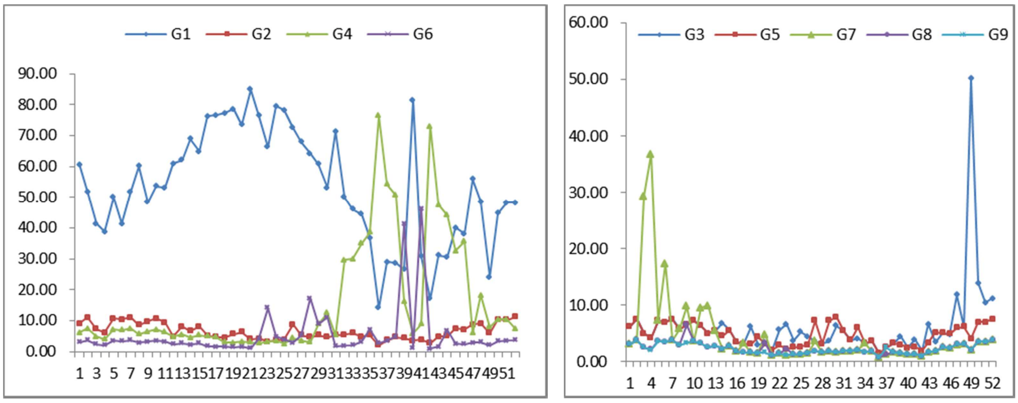
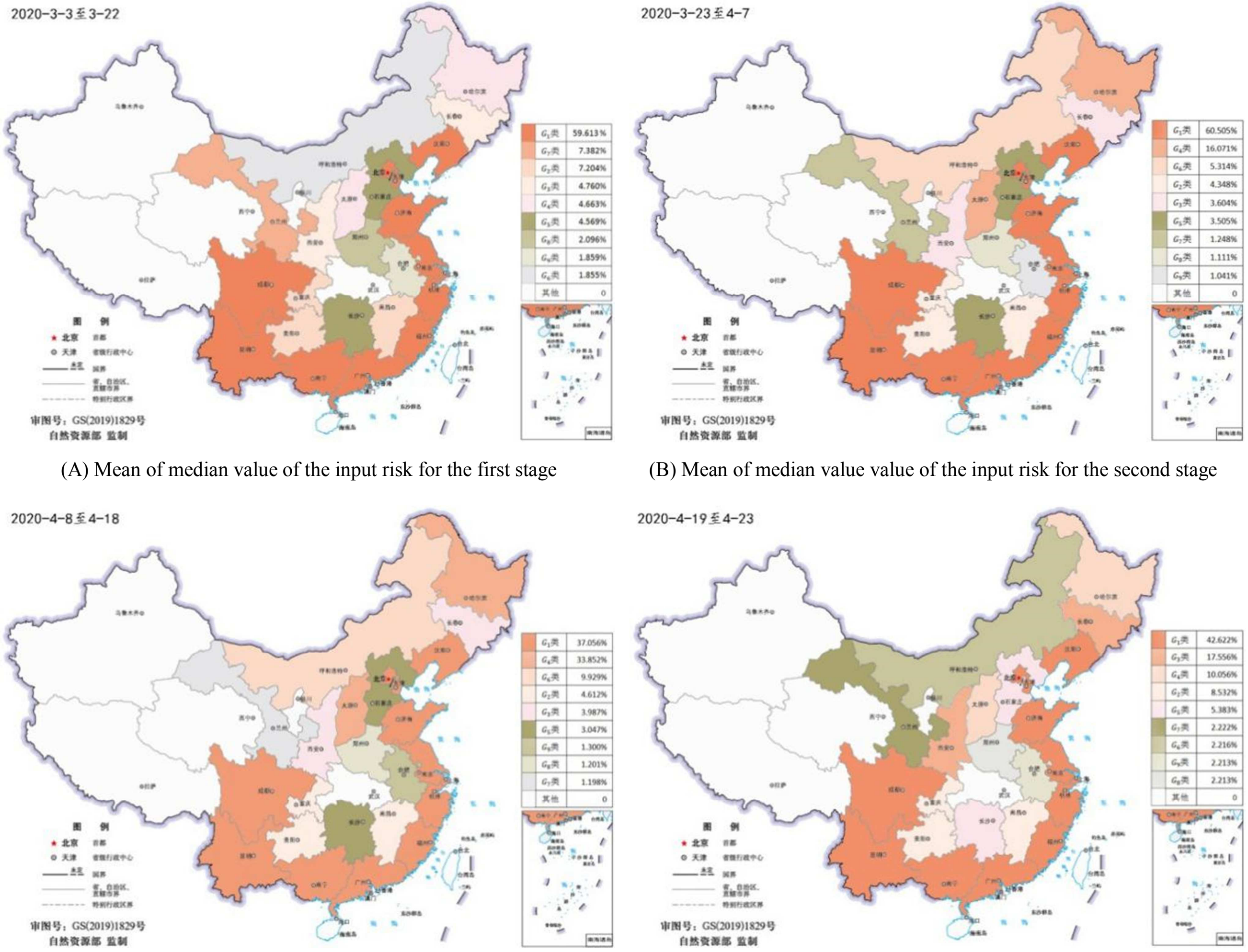


Figure 4. *Risk probability of imported cases from 9 categories of provinces (cities) from March 3 to April 8, 2020 (%).*

Table 3. *Relevant statistics of overseas input risk in 9 categories of provinces (cities) from March 3 to April 8, 2020 (%).*

	G1	G2	G3	G4	G5	G6	G7	G8	G9
Maximum	85.060	10.190	50.040	76.450	7.337	46.200	36.560	5.557	2.658
Minimum	14.020	1.817	1.138	2.172	1.142	0.634	0.476	0.473	0.467
Range	71.040	8.373	48.902	74.278	6.195	45.566	36.084	5.084	2.191
Sd	19.363	2.384	9.067	20.304	1.718	10.144	7.123	0.990	0.624
Mean	53.482	5.905	5.471	14.867	3.998	4.662	3.690	1.615	1.523



(A) Mean of median value of the input risk for the first stage

(B) Mean of median value value of the input risk for the second stage

(C) Mean of median value of the input risk for the third stage

(D) Mean of median value of the input risk for the forth stage

Note: The original map data is a public map of the Ministry of Natural Resources (GS (2019) 1829)

Figure 5. *Risk map of overseas input from March 3 to April 23, 2020.*

In order to intuitively display these characteristics on the map, to calculate the average of the probability estimates (median values) of overseas input risks in four periods as the representative values of this period (see Figure 3), and use Adobe Illustrator software to draw the time-space change chart of overseas input risks (see Figure 4). The comprehensive analysis combined with the input risk estimates (Figure 3) shows that: (1) the input frequency is 0, and there is still overseas input risk. (2) In the long run, the highest risk area is still the first category of 12 provinces and cities represented by Beijing, Shanghai and Guangdong, including 10 coastal/border provinces and cities. There are 7 days in the study period, and the risk value of the fourth category (Heilongjiang and Shanxi) is higher than that of the first category. (2) Before March 21, 2020 (the first stage), the first type of overseas input risk is the highest (59.613%). (3) From March 22 to April 18, 2020 (the second and third stages), the top two overseas input risks are both Category I and Category IV. However, the first category gradually decreased (from 60.50% to 37.00%), and the fourth category gradually increased (from 16.1% to 33.8%). (4) After April 19, 2020, the highest input risks are the first category, the third category (Shaanxi, Jilin), and the fourth category. The overseas input risk of other six categories are less than 5%. For the sixth, seventh, and eighth category, their overseas input risks of are stable around 2%. (5) The last two categories of overseas input risks are the eighth category (Henan) and ninth category (Anhui). (6) The fourth category has the largest fluctuation in overseas input risk.

5. Conclusion and Prospect

Using two clustering methods (Average) and similar (Mcquitty), the 25 provinces and cities with overseas inputs were clustered with indicators, resulting in nine distinct categories. The 12 provinces and cities represented by Beijing, Shanghai and Guangdong are the first category, accounting for 57.42% of the total number of imported cases from abroad. The fourth category Heilongjiang and Shanxi have 12 days of new overseas imported cases more than 11 people per day, and the total proportion of overseas imported cases are 27.86%. Shaanxi and Jilin are the third category, Inner Mongolia is the sixth category, and the total proportion of imported cases abroad are 3.07% and 2.95% respectively. there are more than 98.69% of foreign imported cases in 52 days in above-mentioned 4 categories, 17 provinces and cities. Based on the clustering results, a multinomial distribution under the Bayesian framework is established. The model parameters represent the probability of overseas input risk of 9 categories every day within 52 days. The MCMC method is used to obtain the MC chain of 468 parameters, and the limit distribution is their own posterior distribution, which is suitable for statistical inference with the corresponding MC chain as posterior sample. During the study period, 25 provinces and cities have input risks. The highest risk zone of overseas input is still the first category of 12 provinces and cities represented by Beijing, Shanghai and Guangdong, including 10 coastal/border provinces and cities. During the study period, overseas input risk of the fourth category (Heilongjiang, Shanxi) is higher than the first category, and the input risk of this category fluctuates the most. With 2020-3-22, 4-7, 4-18 as the time nodes, the study period is divided into four stages. The higher risk of overseas input in the four stages is as follows: the first category in the first stage (59.613%); In the second and third stages, the first (from 60.51% to 37.06%), and the fourth category (from 16.07% to 33.85%); In the fourth stage, category I (42.62%), category III (17.56%) and category IV (10.06%). The fourth category has the largest fluctuation in overseas input risk. During the study period, overseas input risks of the eighth (Henan) and ninth (Anhui) categories have been in the last two places.

For overseas input case data with zero expansion, regional loss and large difference in data size. Bayesian multinomial distribution model based on clustering results achieves the estimation of overseas input risk, which is superior to the traditional frequency characterization method. The relevant research needs to be continued by colleagues.

Funding

This research was funded by“Analysis of the Influence Mechanism of Modern Service Industry in Yunnan Province Base on Bayes Method”on the Project of Yunnan University Joint Fund. (2017FH001-068); Supported by the Scientific Research Foundation of Yunnan Provincial Department of Education (2023J1077).

References

[1] Novel coronavirus pneumonia and control Special expert group of the chinese preventive medicine association. An update on the epidemiological characteristics of novel coronavirus pneumonia (COVID-19) [J]. Chinese Journal of Epidemiology, 2020, 4 (02): 139-144.

[2] Epidemiology group of emergency response mechansim of the COVID-19. Chinese center for disease control and prevention. Analysis of epidemiological characteristics of the COVID-19 [J]. Chinese Journal of Epidemiology, 2020, 4 (02): 145-151.

[3] Biao Tang, Xia Wang, Qian Li. Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions [J]. Journal of Clinical Medicine, 2020. doi: 10.3390/jcm9020462.

[4] Shengli Cao, peihua Feng, pengpeng Shi. Aplication of the modified SEIR epidemic dynamics model to prediction and evaluation of the 2019 coronavirus epidemic [J]. Journal of Zhejiang University (Medical Edition). 2020. DOI: 10.3785/j.issn.1008-9292.2020.02.05.

[5] Zifeng Yang, Zhiqi Zeng, Ke Wang, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions [J]. Journal of Thoracic Disease. doi: 10.21037/jtd.2020.02.64.

[6] Domenico Benvenuto, Marta Giovanetti, Lazzaro Vassallo, et al. Application of the ARIMA model on the COVID-2019 epidemic dataset [J], Data in brief, 2020. doi: 10.1016/j.dib.2020.105340.

[7] Adam J Kucharski, Timothy W Russell, Charlie Diamond, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study [OL/EB]. doi: /10.1101/2020.01.31.2001990.

[8] Joseph T Wu, Kathy Leung, Gabriel M Leung, Nowcasting and forecasting the potential domestic and international spread of the COVID-19 outbreak originating in Wuhan, China: a modelling study [J]. The Lancet. 2020, 395 (10225): 689-697.

[9] Duanbing Chen, Wei Bai, Yan Wang, Min Wang, Wuping Yu, Tao Zhou. Quantitative evaluation of prevention and control effect of new coronavirus pneumonia [J/OL]. Journal of University of Electronic Science and Technology of China: 1-6. (31 March 2020). http://kns.cnki.net/kcms/detail/51.1207.T.20200330.1149.002.html.

[10] Shravan Vasishth. Using approximate Bayesian computation for estimating parameters in the cue-based retrieval model of sentence processing [J]. MethodsX, 2020. DOI: 10.1016/j.mex.2020.100850.

[11] Ke Han, Wangping Jia, Wenzhe Cao, Shengshu Wang, et al. Estimation of the real-time basic reproduction number of New Coronavirus pneumonia and evaluation of the current epidemic situation in first tier cities [J/OL]. Journal of the People's Liberation Army Medical College: 1-6 (23 April 2020). http://kns.cnki.net/kcms/detail/10.1117.r.20200421.1109.004.html.

[12] Ensheng Dong, Hongru Du, Gardner Lauren. An interactive web-based dashboard to track COVID-19 in real time [J]. The Lancet. Infectious diseases, 2020. DOI: 10.1016/S1473-3099(20)30120-1.

[13] Weijie Guan, Zhengyi Ni, Yu Hu, et al. Clinical characteristics of coronavirus disease 2019 in China [J]. The New England Journal of Medicine. 2020. DOI: 10.1056/NEJMoa2002032.

[14] Cuifang Qi, liren Yang, Zixuan Yang, Li Shang, et al. Factors affecting the provincial transmission and development of novel Coronavirus pneumonia: Based on data from 30 provinces and cities [J/OL]. Journal of xi'an Jiaotong University (Medical Edition): 1-13 (23 April 2020). http://kns.cnki.net/kcms/detail/61.1399.r.20200417.1413.002.html.

[15] Dali Yi, Gaoming Li, Huiming Leng. Cluster analysis of regional differences in the development of novel Coronavirus pneumonia [J/OL]. Journal of Chongqing Medical University: 1-6 (23 April 2020). https://doi.org/10.13406/j.cnki.cyxb.002386.

[16] Zhang Liu, Jiale Qian, Yunyan Du, et al. Multi-level spatial distribution estimation model of the interregional migrant population using multi-source spatio-temporal big data: Take population emigrated from Wuhan during the COVID-19 epidemic as an example [J]. Journal of Geoinformation Science, 2020, 22 (02): 147-160.

[17] Mingxiang Feng, Zhixiang Fang, Xiongbo Lu, et al. Estimation method of temporal and spatial spread of novel coronavirus pneumonia on, the scale of traffic analysis area: A case study of wuhan city [J/OL]. Journal of Wuhan University (Information Science Edition): 1-12 (23 April 2020). https://doi.org/10.13203/j.whugis20200141.

[18] Xiaoqun He. Multivariate statistical analysis (5th Edition) [M]. Beijing: China Renmin University Press, 2016, 1, 4: 52-60.

[19] Yi Xue. Statistical modeling and R software (1st Edition) [M]. Beijing: Tsinghua University press, 2007, 4: 402-418.

[20] Nguyen X. Borrowing strengh in hierarchical bayes: Posterior concentration of the dirichlet base measure [J]. Bernoulli, 2016, 22 (3): 1535-1571.

[21] Radford M Neal. Suppressing random walks in markov chain monte carlo using ordered overrelaxation [M]. Springer Netherlands, 1998. 205-230.

[22] Lunn D, Jackson Christopher, et al. A Practical Introduction to Bayesian Analysis [M]. CRC Press, 2013.

[23] Shisong Mao, Yiming Cheng, et al. Probability theory and mathematical statistics (2nd Edition) [M]. Beijing: Higher Education Press, 2011: 332-338.