Using Social Media to Mine and Analyse Public Opinion Related to COVID-19 in China

1. Introduction

As of March 7, 2020, the global number of confirmed cases of novel coronavirus disease (COVID-19) surpassed 100,000, covering more than 100 countries (World Health Organization, 2020a). As a growing number of confirmed cases of infections was reported, the Chinese government took prompt response measures to combat the virus. On January 9, 2020, the causative agent of this pneumonia was initially confirmed as a novel coronavirus. On January 20, the National Health Commission of the People's Republic of China (NHFPC) classified COVID-19 as a category B infectious disease based on the Law on Prevention and Control of Infectious Diseases and took preventive and control measures for category A infectious diseases. On January 23, Wuhan City was put under lockdown to contain the outbreak. On February 2, Huoshenshan Hospital in Wuhan, a 1,000-bed makeshift hospital for treating infected patients, was built just in 10 days. On February 8, China completed work on Leishenshan Hospital, another 1,600-bed makeshift hospital in Wuhan. As of February 10, the number of confirmed cases of COVID-19 in China surpassed 42,000. In the early stages of the COVID-19 outbreak in China, there was a fierce race between the growth of patients and the allocation of medical resources.

Coinciding with the Lunar New Year Festival, COVID-19 not only disrupted people's normal lives, it also attracted great attention from all circles of society. In these circumstances, billions of people eagerly acquired information about COVID-19 through social media. Topics and sentiments related to COVID-19 spread rapidly, thus influencing public behaviour during the epidemic. Analyses of public opinion are important for improving emergency responses, enhancing sentiment awareness, and supporting decision making.

This study aimed to identify public opinion during the COVID-19 outbreak from social media and analyse its spatial-temporal characteristics from January 9 to February 10, 2020 in China. A topic extraction and classification model was built to identify the topics of COVID-19-related Weibo and uncover public sentiments in response to COVID-19. A series of topics and temporal and spatial distributions were identified and discussed.

2. Data and Methods

2.1. Data and Data Pre-Processing

Sina-Weibo (http://us.Weibo.com), often referred to as Weibo, is one of the most popular social media platforms in China. Weibo had over 516 million active users each month in 2019. This study acquired Weibo texts related to COVID-19. Using Weibo Application Programming Interfaces (APIs), Weibo messages related to COVID-19 were collected with ‘pneumonia’ and ‘coronavirus’ as the keyword with timestamps between 00:00 on January 9, 2020 and 24:00 on February 10, 2020. The following information was extracted: user ID, timestamp (i.e., the time at which the message was posted), text (i.e., the text message posted by a user), and location information.

The original Weibo texts contain interfering information such as http hyperlinks, spaces, punctuation marks, hashtags, and @users. Text filtering was thus necessary to eliminate noise and improve the efficiency of word segmentation. These types of interfering information were removed by regular expression operations (‘re’ module) in Python. Very short Weibo texts (less than four words) and duplicated Weibo texts were deleted. That left 1,413, 297 Weibo messages, including 105, 330 texts with geographical location information.

2.2. Method

A topic extraction and classification model combining the LDA model and the random forest (RF) algorithm was used to hierarchically process COVID-19-related Weibo texts. The first step was to mine and generalize the topics from the COVID-19-related Weibo sample using the LDA model. Then, topic extraction results were utilized as training samples for the RF algorithm to classify the Weibo data. As shown in Figure 1, the COVID-19-related Weibo were generalized into seven topics: ‘events notification’, ‘popularization of prevention and treatment’, ‘government response’, ‘personal response’, ‘opinion and sentiments’, ‘seeking help’, and ‘making donations’. A secondary classification was implemented to divide ‘personal response’, ‘opinion and sentiments’, and ‘seeking help’ into thirteen more detailed sub-topics, including ‘fear and worry’, ‘questioning the government and media’, ‘condemning bad habits’, ‘objective comment’, ‘taking scientific protective measures’, ‘blessing and praying’, ‘appealing for aiding patients’, ‘willing to return work’, ‘staying at home and taking necessary precautions’, ‘popularizing anti-epidemic knowledge in family’, ‘seeking medical help’, ‘seeking relief materials’, and ‘other’.

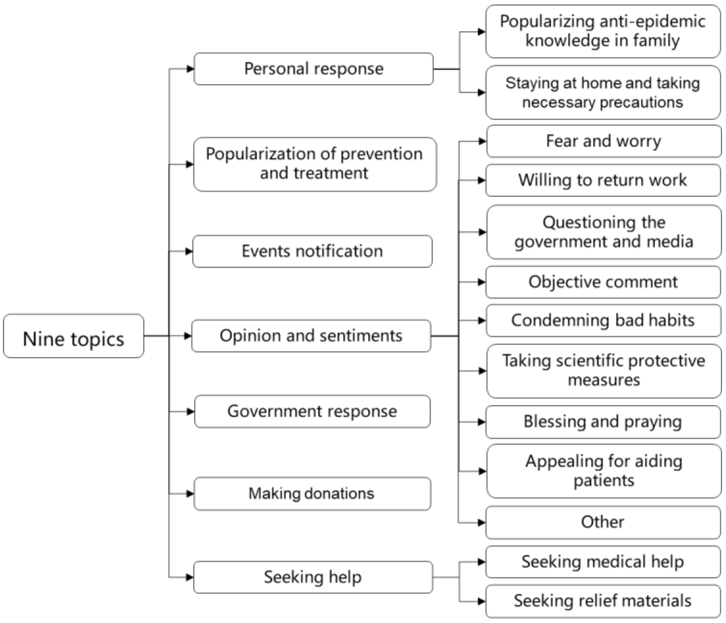


Figure 1. The classification of COVID-19 related topics and sub-topics.

The processes of topic extraction and classification are shown in Figure 2, including the steps of word segmentation and topic extraction using the LDA and RF models.

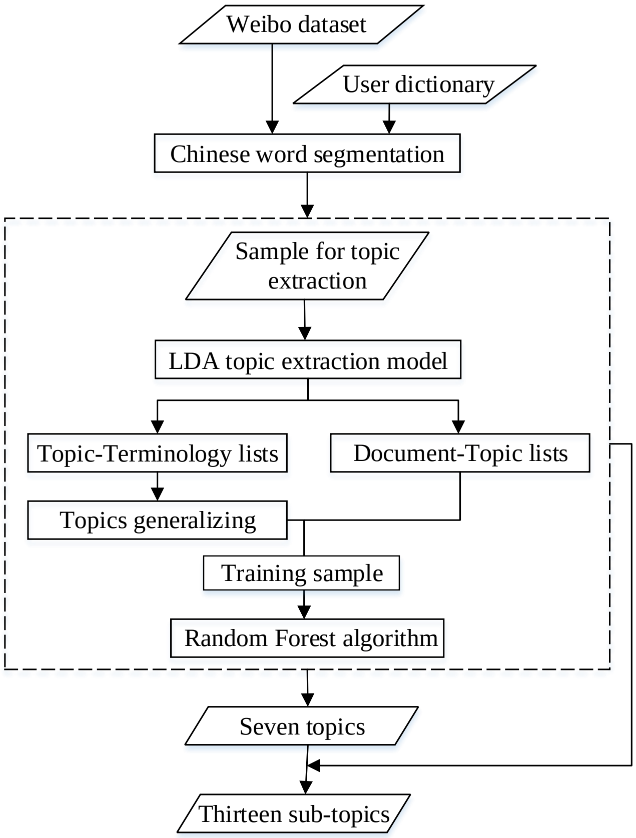


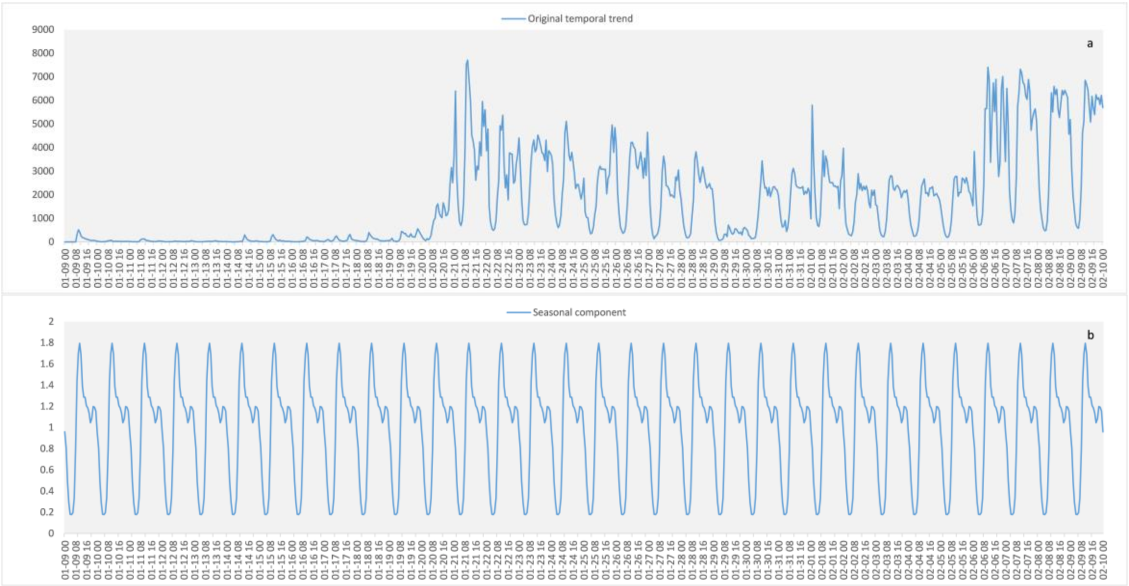
Figure 2. The processes of topic extraction and classification.

3. Results

3.1. Spatial-Temporal Analysis

3.1.1 Time series analysis

The results of the time series analysis of COVID-19-related Weibo texts are shown in Figure 3. Split by day, Figure 3a shows that the lowest point of the Weibo number on the curve for each day appeared at 04:00, after which the curve began to rise sharply. Figure 3b shows the lowest point of cyclical change occurring at 06:00 every day, with two daily peaks around 11:00 and 23:00. Figure 3c shows the seasonally adjusted time series, which shows the trend of the number of COVID-19-related Weibo after eliminating the seasonal factor. Figure 3d shows the trend component reflecting the trends of the number of COVID-19 related Weibo. After COVID-19 occurred, a slight increase appeared for a short time, then the amount increased sharply on January 20. The fluctuation reached a peak on January 21, and then began to decrease but fluctuated until January 29. The curve rose obviously on January 31 and reached a peak on February 1. It then steadily fluctuated from February 2 to 5, started to climb on February 6, and then steadily declined after reaching the highest peak on February 7.



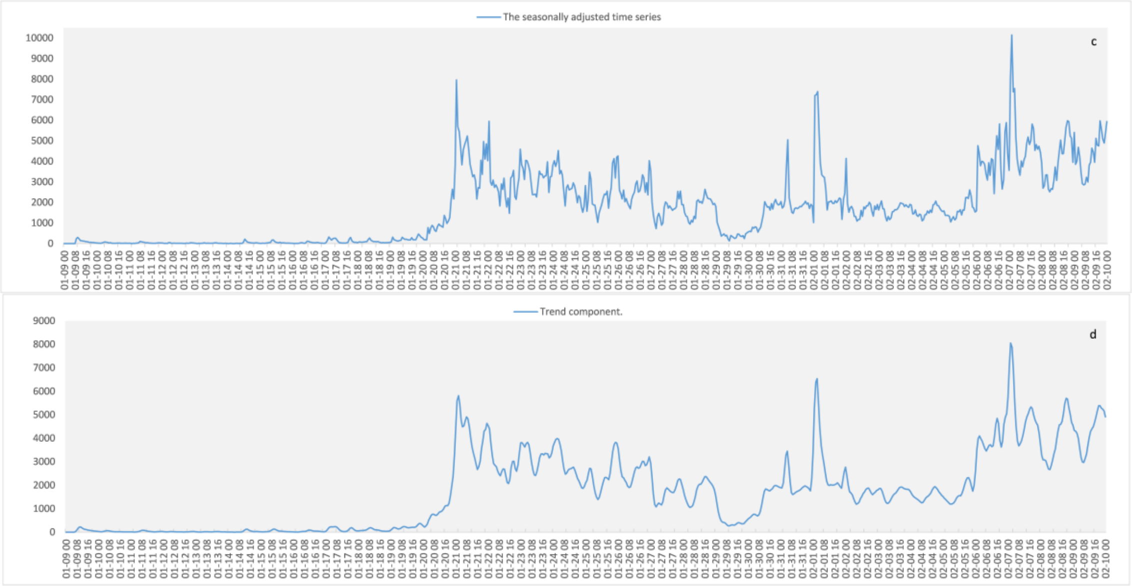
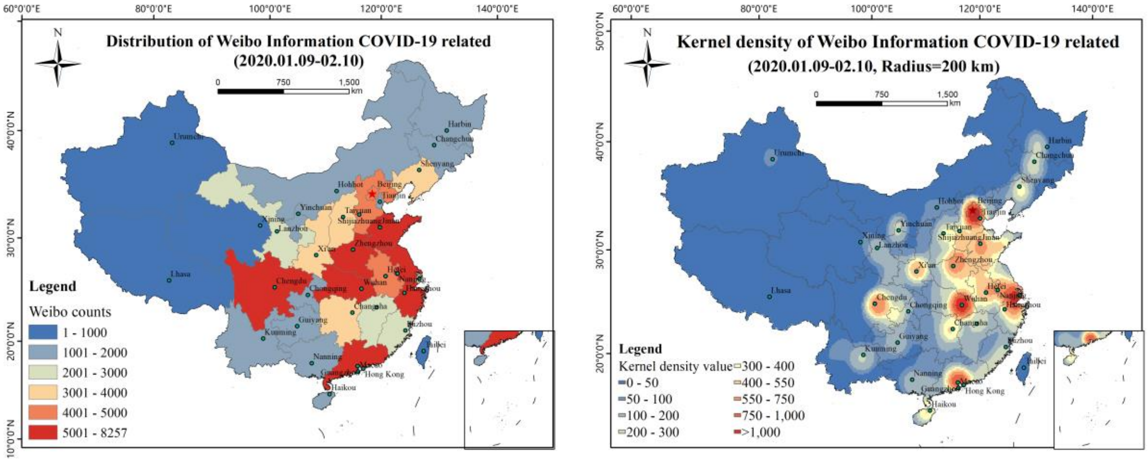


Figure 3. The seasonal trend decomposition of the temporal trends of COVID-19-related Weibo. (a) Original temporal series; (b) seasonal component; (c) seasonally adjusted time series; (d) trend component.

3.1.2 Spatial analysis

The spatial distribution of Weibo related to COVID-19 is shown in Figure 4. The Weibo numbers were mainly concentrated in the east-central parts of China, as shown in Figure 4a. There were more than 5,000 Weibo texts in Shandong Province (the capital of Jinan), Hubei Province (capital of Wuhan), Henan (capital of Zhengzhou), Guangdong (capital of Guangzhou), Sichuan (the capital of Chengdu), and Jiangsu (capital of Nanjing), Anhui Province (capital of Hefei), Hebei Province (capital of Shijiazhuang), Beijing, Shaanxi (capital of Xi’an), Liaoning (capital of Shenyang), Hunan (capital of Changsha), and Shanxi (capital of Taiyuan). Figure 4b shows the spatial distribution of the kernel density with a search radius of 100 km, indicating that the high-density areas of Weibo related to COVID-19 were in Wuhan, Beijing, Shanghai, Guangzhou, Chengdu, Xi’an, and Zhengzhou, and presents a continuous trend among the hot points of Wuhan, Beijing, and Shanghai.



(a) (b)

Figure 4 The spatial distribution of Weibo related to COVID-19.

3.2. Topic Analysis

3.2.1. Topic description

Figure 5 illustrates the statistical results of the percentage of first-level topics of COVID-19. ‘Opinion and sentiments’ accounted for 34.42% of all topics. ‘Popularization of prevention and treatment’ and ‘government response’ were the second and third most frequent, at 18.97% and 16.29%, respectively. The proportion of ‘events notification’ and ‘personal response’ comprised 13.94% and 12.82%, respectively. ‘Seeking help’ and ‘making donations’ then accounted for 2.01% and 1.55%, respectively.

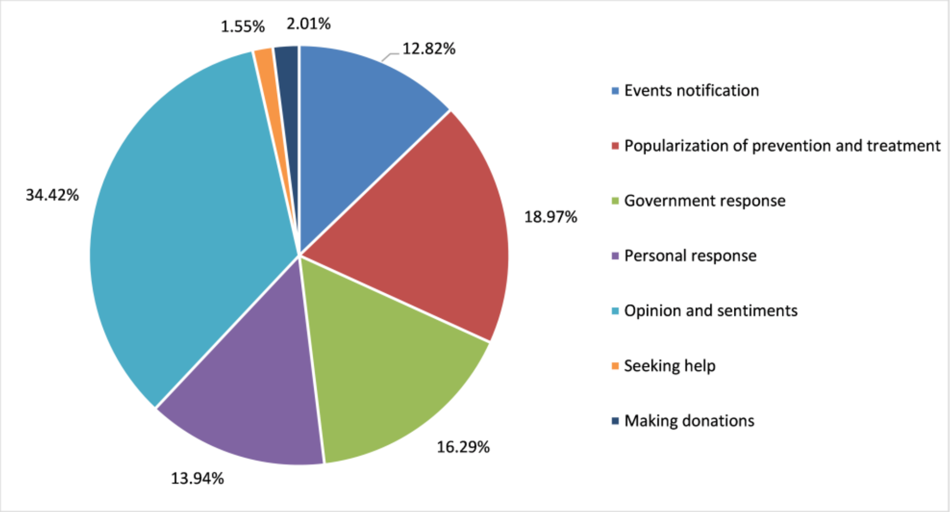


Figure 5. Classification of topics in Weibo texts related to COVID-19

A more in-depth analysis of the proportions of sub-topics is presented in Figure 6. ‘Staying at home and taking necessary precautions’, ‘blessing and praying’, and ‘objective comment’ were the three most widespread sub-topics, accounting for 23.26%, 20.89%, and 14.99% of texts. The proportion of ‘taking scientific protective measures’ and ‘fear and worry’ comprised 12.48% and 10.47%, respectively. This was followed by ‘condemning bad habits’ and ‘seeking medical help’, which accounted for 6.02% and 4.14%. The proportion of other sub-topics was less than 3%.

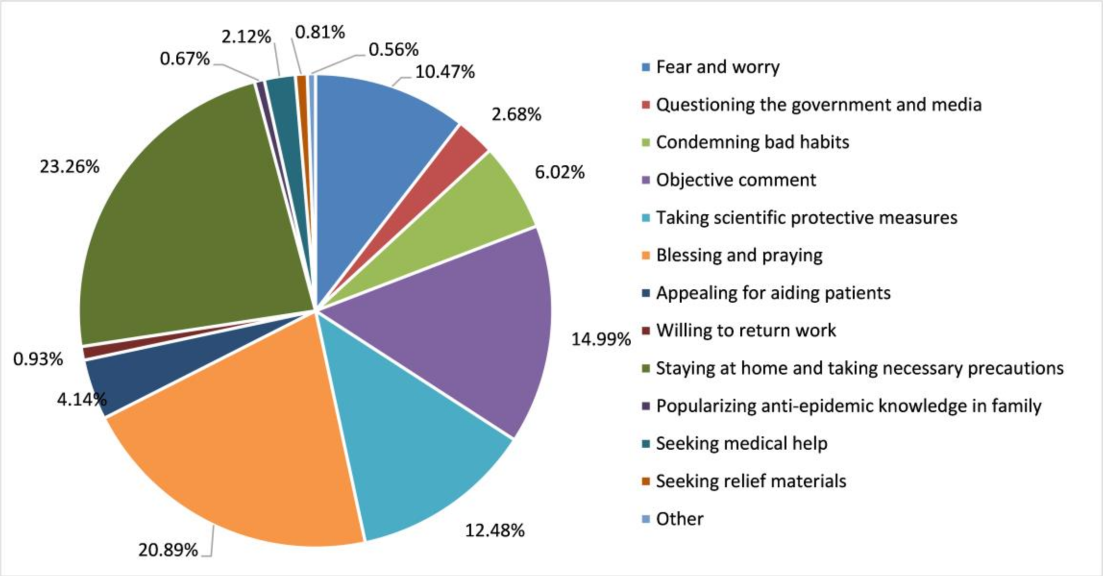


Figure 6. Classification statistics of sub-topics in Weibo texts related to COVID-19

After computing precision, recall, and F1-measure values, the classification accuracy of the topics and sentiments was presented in Table 1. For the seven topics, the precision was found to be 83% and F1 was 82%. For the thirteen sub-topics, the precision and F1 values were 78% and 76%, respectively.

Table 1. Evaluation results of topic classification.

	Topics	**Sub-topics**
Precision (P)	83%	78%
Recall (R)	82%	77%
F1	82%	76%

3.2.2. Temporal Trend of Topics

To display accurate temporal changes in the different topics, the number of Weibo texts for each topic was counted using one-hour time intervals as shown in Figure 7. The topics of ‘events notification’, ‘popularization of prevention and treatment’, ‘personal response’, and ‘opinion and sentiments’ all climbed from January 19, reaching a peak on the 21st. The curve then steadily declined towards the 29th and rose slowly to February 1. There was a small peak on February 1, then it stabilized and reached a peak again on February 5. The topics of ‘government response’ and ‘making donations’ started to rise steadily from January 20, then declined after showing a small peak around January 26, after which it started to climb on February 4 and reached a peak on February 5. ‘Seeking help’ started to rise suddenly on January 22, showing a small peak before and after Wuhan was placed under lockdown on the 23rd, then climbing on February 4, reaching a peak on February 6, and then levelling off.

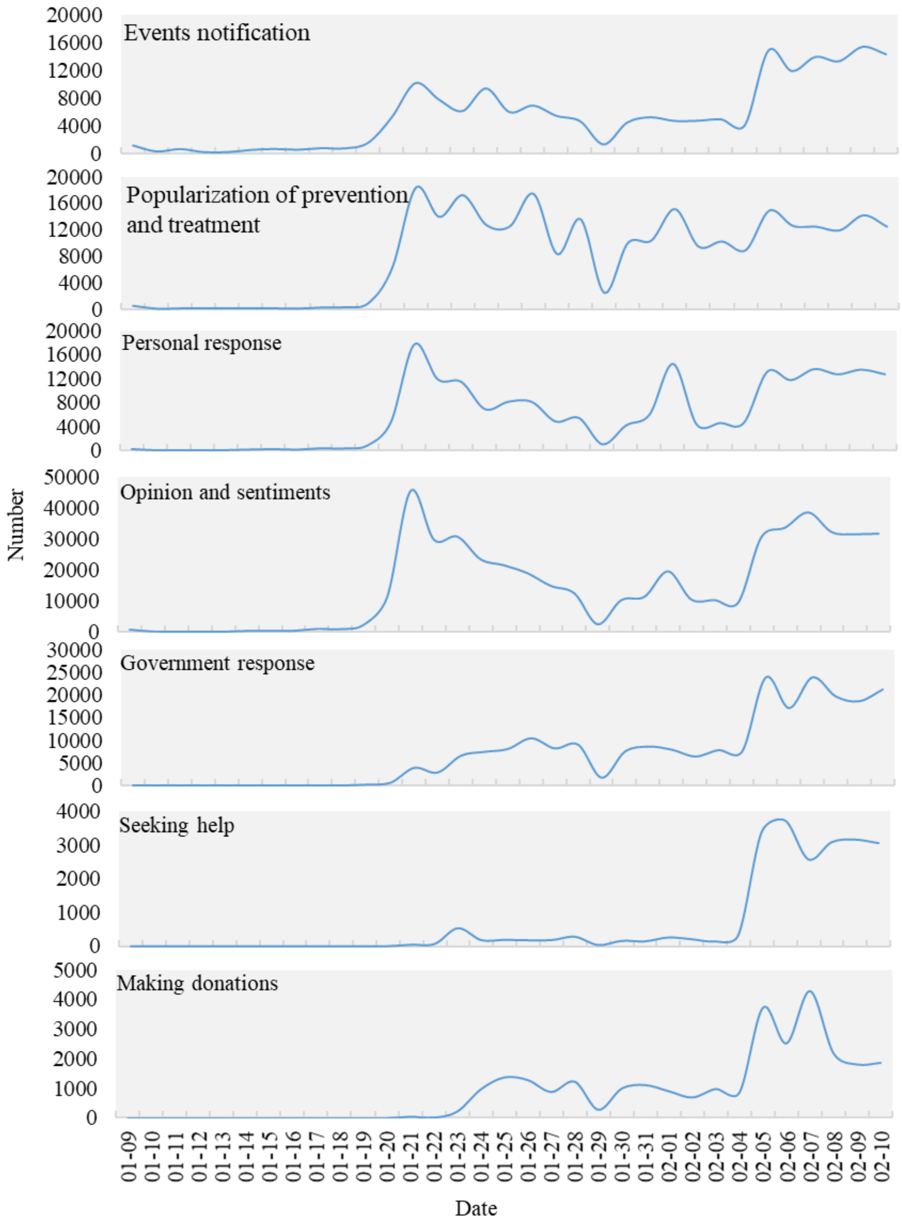


Figure 7. The temporal series of topics during COVID-19

Figure 8 presents the time series of all sub-topics except ‘other’. From a perspective of the general trends, the three sub-topics, ‘questioning the government and media’, ‘staying at home and taking necessary precautions’, and ‘taking scientific protective measures’ showed a similar variation tendency over time. The numbers of texts on those three sub-topics improved quickly on January 20 and peaked on the 21st, then gradually decreased but fluctuated towards January 29, rose obviously on January 31, and reached a peak on February 1. Since then, the curve has been steadily fluctuating, beginning to rise on February 5. ‘Fear and worry’, ‘objective comment’, and ‘blessing and praying’ climbed from January 20, reached their peak on 21, fell steadily, then rose again on February 5 and stabilized. ‘Appealing for aiding patients’ and ‘seeking medical help’ suddenly increased from February 6 and reached a peak around February 8. After that, the ‘appealing for aiding patients’ showed a downward trend, and the ‘seeking medical help’ remained a high concern. ‘Popularizing anti-epidemic knowledge in family’ and ‘condemning bad habits’ both started to climb on January 20. After reaching a summit on the 21st, the decline since stabilized. ‘Seeking relief materials’ began to rise on January 22, fell to a peak on 23, then stabilized after rising on February 5. ‘Willing to return work’ had a slight increase and fluctuation since January 20 and has shown a significant upward trend since February 4.

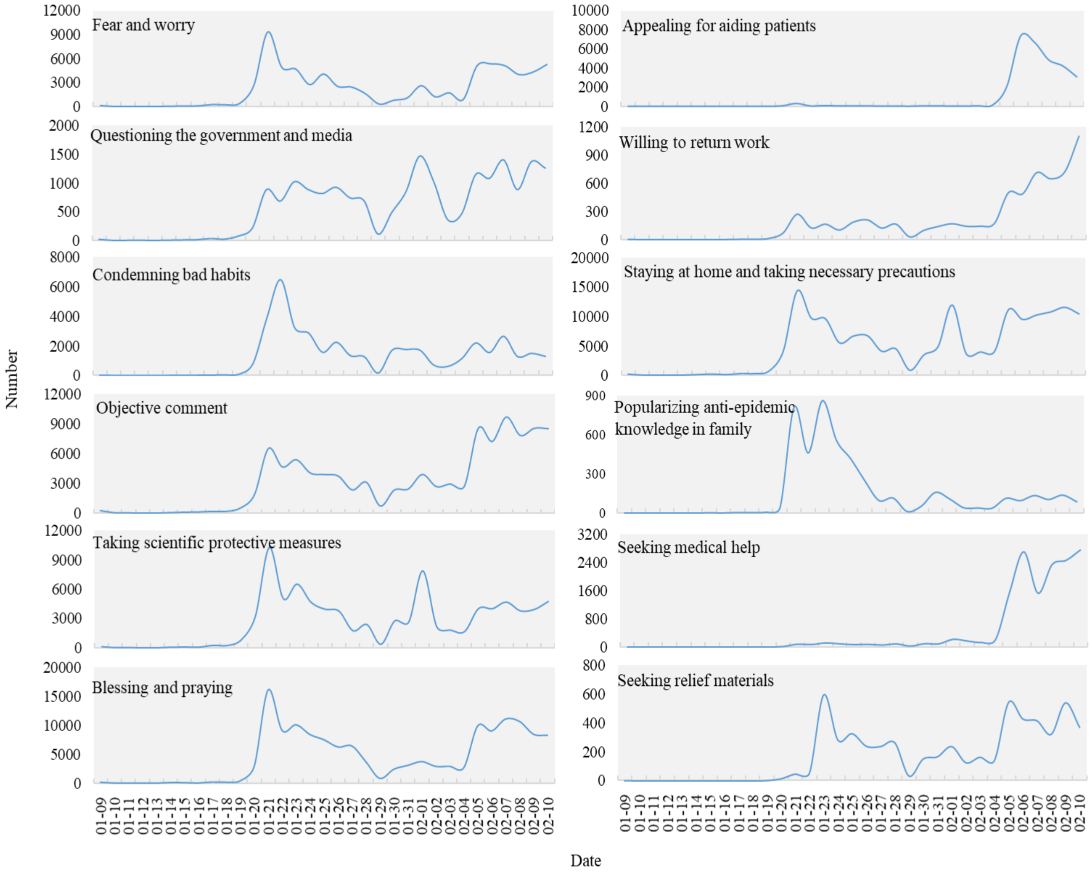


Figure 8. The temporal series of sub-topics during COVID-19

3.2.3. Spatial Distribution of Topics

Kernel density analysis (radius of 200 km) was carried out on Weibo with geographical locations in each topic, as shown in Figure 9. The spatial distribution of ‘events notification’, ‘popularization of prevention and treatment’, ‘government response’, ‘personal response’ and ‘opinion and sentiments’ is similar to the general characteristics of Figure 4b, forming hot spots in Beijing-Tianjin-Hebei, Shandong, Henan, Hubei, Yangtze River Delta, Sichuan, and Guangdong, but there are differences within the topics. ‘Events notification’ takes Beijing, Wuhan, Shanghai, and Sichuan as prominent high-value areas, and the areas of the Beijing-Tianjin-Hebei cross border area, east Hubei, and the Jiangsu-Zhejiang-Shanghai cross border areas are the main nodes in a continuous pattern. ‘Popularization of prevention and treatment’ is presented with Beijing, Guangzhou, and Shanghai as the prominent high values, supplemented by Wuhan, Chengdu, Hefei, Zhengzhou, and other high-value areas. ‘Government response’ has Beijing, Sichuan, and Xi'an as high values, though Zhengzhou, Wuhan, Changsha, Shanghai, Guangzhou, Haikou, and other cities have responded significantly. ‘Personal response’ is prominently reflected by Beijing, Shanghai, Guangzhou, and Wuhan, with Beijing, Wuhan, and Shanghai as the centre and Guangzhou and Chengdu as relatively independent high-value areas. ‘Opinion and sentiments’ was more prominent in high-value areas around Wuhan, followed by the Yangtze River Delta, Beijing-Tianjin-Hebei, and the Pearl River Delta urban agglomeration. ‘Seeking help’ and ‘making donations’ show totally different characteristics. ‘Seeking help’ appears significantly around Wuhan and shows a trend of diffusion to the surrounding areas, especially to the north. ‘Making donations’ has Beijing and Hainan as high values and spreads across the country, but is relatively concentrated in urban areas around Wuhan, the Yangtze River Delta region, Chengdu-Chongqing region, Guangzhou, Zhengzhou, and even Haikou in the south.

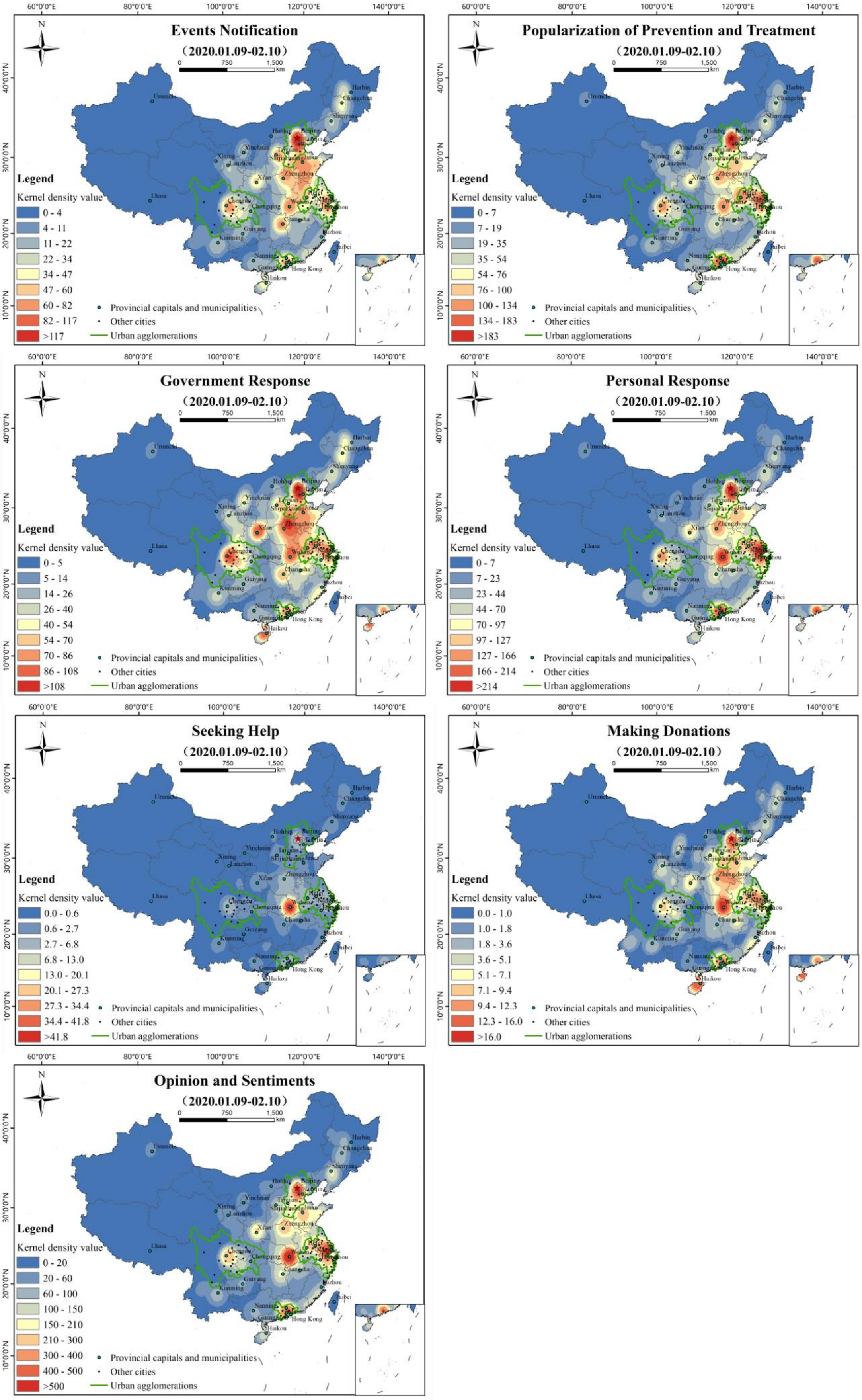
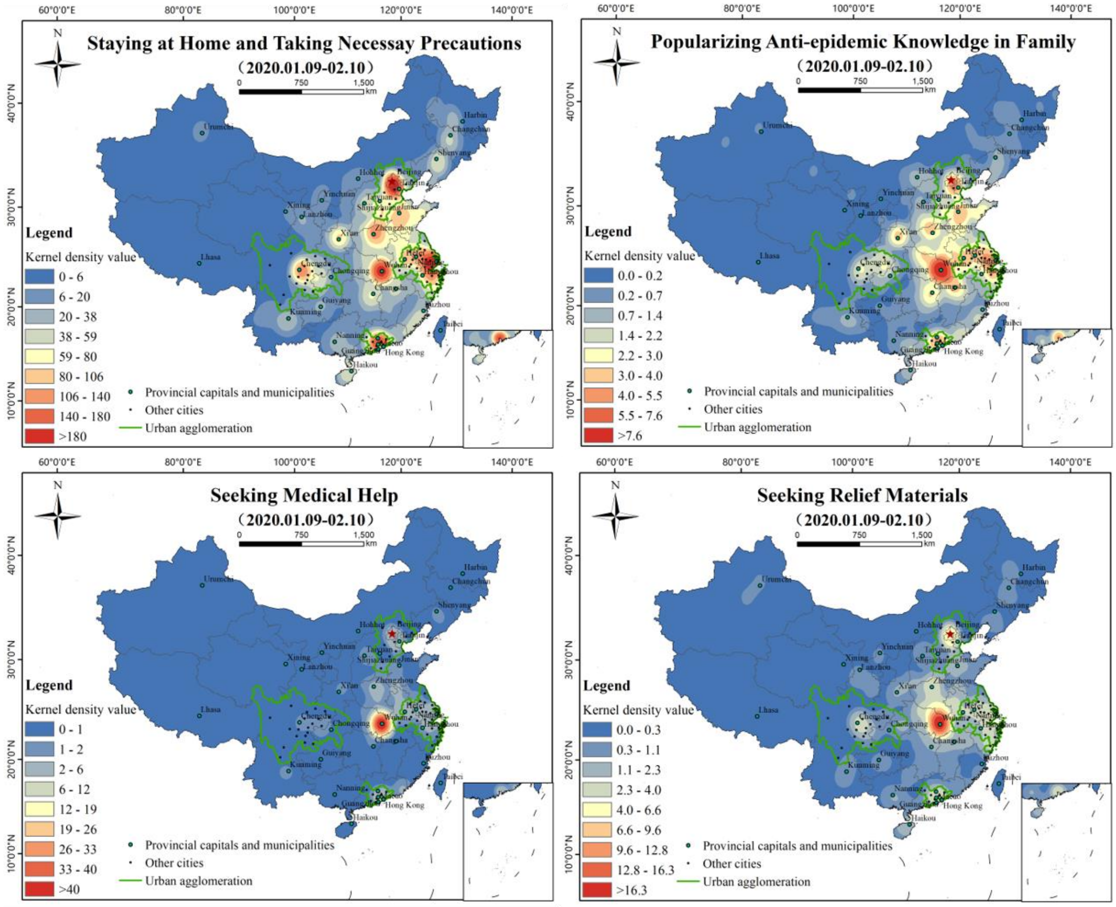


Figure 9. Kernel density analysis of each primary topic (search radius = 200 km)

The spatial distributions of the kernel density estimation of the thirteen sub-topics are shown in Figure 10. Except for ‘appealing for aiding patients’, ‘seeking medical help’, and ‘seeking relief materials’, the spatial distribution of most topics is similar to the general characteristics of Figure 4b. ‘Fear and worry’ formed high-value areas in Wuhan, Shanghai, Suzhou, Jiaxing, and other cities. ‘Questioning the government and media’ is mainly reflected in Wuhan, supplemented by the Beijing-Tianjin-Hebei transboundary area, east Hubei, the Jiangsu-Zhejiang-Shanghai neighbourhood area, and Guangzhou and Chengdu, two relatively independent high-value areas. ‘Condemning bad habits’ is distributed in dots as a whole. Beijing is a high-value region with prominent dots, and east Hubei, the Jiangsu-Zhejiang-Shanghai cross border area, Guangzhou, and Wuhan are independent high-value regions. ‘Objective comment’ takes Wuhan as a prominent high-value area, supplemented by Beijing, Shanghai, Guangzhou, and other high-value areas. ‘Taking scientific protective measures’ is a prominent spot-shaped high-value area in Beijing, Wuhan, and Shanghai, and the areas within the Beijing-Tianjin-Hebei neighbourhood area, east Hubei, the Jiangsu-Zhejiang-Shanghai transboundary areas are the main nodes in a continuous pattern. ‘Blessing and praying’ is centred on the contiguous areas of Beijing, Wuhan, and Shanghai, while Guangzhou, Chengdu, and Zhengzhou are relatively independent high-value areas. ‘Appeal for aiding patients’ takes Wuhan as the centre of east Hubei as the high-value area, and Beijing, Shanghai, and the neighbourhood area as relatively high-value areas. ‘Willing to return work’ shows that Beijing, Guangzhou, and Shanghai are prominent high-value areas, supplemented by Wuhan, Chengdu, Hefei, Jinan, and other relatively high-value areas. ‘Staying at home and taking necessary precautions’ is led by Wuhan, with Beijing, Wuhan, Shanghai, and Guangzhou as the highlighted high-value areas, and the Beijing-Tianjin-Hebei cross border area, east Hubei, the Yangtze river delta, and the Pearl River Delta as the main nodes, showing a continuous trend. ‘Popularizing anti-epidemic knowledge in family’ is concentrated in Wuhan and its surrounding cities, supplemented by relatively high-value areas such as Beijing, Shanghai, and Guangzhou. ‘Seeking medical help’ and ‘seeking relief materials’ are prominently concentrated in Hubei. ‘Seeking medical help’ appears in Wuhan and spreads to the surrounding area, especially to the east. The overall distribution of ‘seeking relief materials’ and ‘seeking medical help’ showed a similar distribution trend, with Wuhan and its surrounding areas as high-value areas. ‘Appealing for aiding patients’ is mainly distributed in Wuhan, Beijing, Shanghai, and other regions.



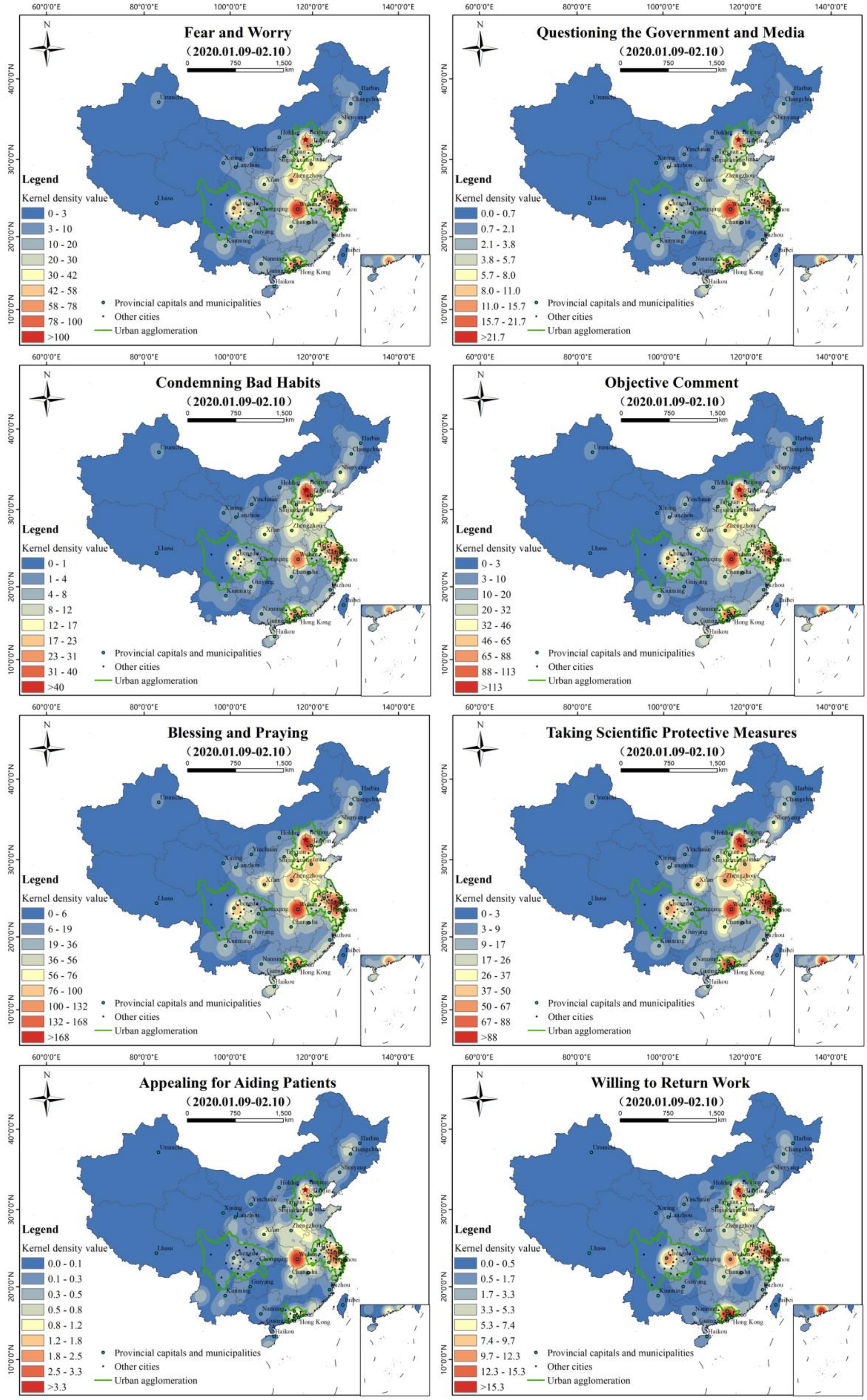


Figure 10. Kernel density analysis of each secondary topic (search radius=200 km)

4. Conclusions

This study comprehensively analysed social media data in the early stage of COVID-19 in China and proposed a topic extraction and classification model. The results of the evaluation show that the approach for topic extraction is accurate and viable for understanding public opinions. We obtained seven topics and thirteen sub-topics related to COVID-19 from Weibo texts and analysed their temporal-spatial distributions. (1) The topic with the most quantities in the early stage of COVID-19 was 'opinion and sentiments' 'staying at home and taking necessary precautions', 'blessing and praying', 'taking scientific protective measures', and 'popularizing anti-epidemic knowledge in family' was the most-expressed sub-topic. This finding indicates that timely release of information from the government was helpful in stabilizing public opinion in the early stage of COVID-19. (2) The temporal changes in Weibo texts are synchronous with the development of the COVID-19 outbreak. The spatial distribution of COVID-19- related Weibo texts shows a distribution pattern that Beijing-Tianjin-Hebei, the Yangtze River Delta, the Pearl River Delta, and the Chengdu-Chongqing urban agglomeration were significant high-value areas besides Wuhan. This means that the temporal-spatial distribution of opinions was related to the severity of the epidemic, the degree of population aggregation, and the level of economic development. (3) The spatial distribution of public opinions is regionally different and has a scale feature. Information on these topics is very important as a reference for emergency response and post-disaster management. It is suggested that the government should strengthen its response to the key epidemic area and the urban agglomerations, and formulate accurate response countermeasures following the public's demands in controlling the crisis.

Nevertheless, this study has some limitations. First, the specific reasons for the temporal-spatial distribution of COVID-19-related Weibo texts need further exploration with more information. Second, the paper only analysed texts from social media, while other content, such as pictures and videos in blogs, may also be informative. With COVID-19, which has been characterized as a pandemic by WHO (World Health Organization, 2020b), we will continually acquire new data from Weibo, train and improve the model, analyse the changes and driving mechanisms of public opinion, and provide active references for governmental responses.

References:

Ignore.