Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Early warning of COVID-19 hotspots using human mobility and web search query data

Takahiro Yabe\textsuperscript{a,b}, Kota Tsubouchi\textsuperscript{c}, Yoshihide Sekimoto\textsuperscript{d}, Satish V. Ukkusuri\textsuperscript{a,*}

\textsuperscript{a} Lyles School of Civil Engineering, Purdue University, 550 Stadium Mall Avenue, West Lafayette, IN 47907, USA
\textsuperscript{b} Institute for Data, Systems, and Society, Massachusetts Institute of Technology, 50 Ames St, Cambridge, MA 02142, USA
\textsuperscript{c} Yahoo Japan Corporation, Kioi Tower, Tokyo, Garden Terrace Kioicho, 1-3, Kioi-cho, Chiyoda-ku, Tokyo, Japan
\textsuperscript{d} Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba Meguro-Ku, Tokyo 153-8505, Japan

\textbf{ARTICLE INFO}

Keywords:
COVID-19
Epidemics
Human mobility
Web search data

\textbf{ABSTRACT}

COVID-19 has disrupted the global economy and well-being of people at an unprecedented scale and magnitude. To contain the disease, an effective early warning system that predicts the locations of outbreaks is of crucial importance. Studies have shown the effectiveness of using large-scale mobility data to monitor the impacts of non-pharmaceutical interventions (e.g., lockdowns) through population density analysis. However, predicting the locations of potential outbreak occurrence is difficult using mobility data alone. Meanwhile, web search queries have been shown to be good predictors of the disease spread. In this study, we utilize a unique dataset of human mobility trajectories (GPS traces) and web search queries with common user identifiers (> 450 K users), to predict COVID-19 hotspot locations beforehand. More specifically, web search query analysis is conducted to identify users with high risk of COVID-19 contraction, and social contact analysis was further performed on the mobility patterns of these users to quantify the risk of an outbreak. Our approach is empirically tested using data collected from users in Tokyo, Japan. We show that by integrating COVID-19 related web search query analytics with social contact networks, we are able to predict COVID-19 hotspot locations 1–2 weeks beforehand, compared to just using social contact indexes or web search data analysis. This study proposes a novel method that can be used in early warning systems for disease outbreak hotspots, which can assist government agencies to prepare effective strategies to prevent further disease spread.

Human mobility data and web search query data linked with common IDs are used to predict COVID-19 outbreaks. High risk social contact index captures both the contact density and COVID-19 contraction risks of individuals. Real world data was collected from 200 K individual users in Tokyo during the COVID-19 pandemic. Experiments showed that the index can be used for microscopic outbreak early warning.

1. Introduction

The coronavirus pandemic (COVID-19) has inflicted significant health and economic impacts across the globe. Due to the contagious nature of the disease, monitoring and controlling how people come in contact with each other has shown to be key in containing the disease (Zhang et al., 2020). Therefore, in response to the COVID-19 pandemic, countries have taken various non-pharmaceutical interventions (NPIs) (e.g., lockdowns, social distancing, testing and tracing) (Flaxman et al., 2020), while facing the conundrum of balancing the trade-off between benefits on public health and depletion of economic performance (Ukkusuri, Yabe, & Seetharam, 2020). In order to monitor and evaluate the impacts of such interventions, large scale mobile phone data has been utilized as an effective data source (Fan & Stewart, 2021; Oliver et al., 2020). Studies on human mobility analysis have used such data to model disease dynamics (Bengtsson et al., 2015; Dodge et al., 2021; Finger et al., 2016; Tizzoni et al., 2014; Wesolowski et al., 2012). During the COVID-19 crisis, various stakeholders have utilized large-scale mobility datasets to evaluate the effects of NPIs in various regions (Bonato et al., 2020; Cintia et al., 2020; Dahlberg et al., 2020; Gao, Kang, Liang, & Kruse, 2020; Klein et al., 2020; Kraemer et al., 2020; Lai et al., 2020; Long & Reuschke, 2021; Pepe et al., 2020; Santana et al., 2020; Schlosser, Maier, Hinrichs, Zachariae, & Brockmann, 2020; Wellenius et al., 2020; Yabe et al., 2020).

* Corresponding author.
E-mail addresses: tyabe@mit.edu (T. Yabe), ktsubouc@yahoo-corp.jp (K. Tsubouchi), sekimoto@iis.u-tokyo.ac.jp (Y. Sekimoto), sukkusur@purdue.edu (S.V. Ukkusuri).

https://doi.org/10.1016/j.compenvurbsys.2021.101747
Received 11 July 2021; Received in revised form 6 December 2021; Accepted 8 December 2021
Available online 15 December 2021
0198-9715/© 2021 Elsevier Ltd. All rights reserved.
The aforementioned studies have shown the effectiveness of using large-scale mobility data to monitor physical co-location of the population (which can be used as a proxy for social contacts) in a fine-grained spatial and temporal scale. More specifically, metrics such as “social contact index (SCI)” have been proposed, and their strong relationships with the estimated transmissibility of COVID-19 (i.e. effective reproduction number $R_0$) has been shown (Yabe et al., 2020). However, such metrics are incapable of predicting outbreaks beforehand, since such analysis can only be conducted in a retrospective manner. This is a critical drawback since countries are already facing second and third waves of COVID-19 (as of October 2020), and countries are in need of effective early warning systems that can predict when and where the next outbreaks would occur beforehand.

To tackle this problem, we utilize a unique dataset that contains both the GPS location data (mobility trajectories) and web search queries of users, which are linked with common user identifiers (IDs). We hypothesize that users who search COVID-19 symptom related queries more frequently and intensely have a higher risk of having contracted the virus. By integrating the mobility analysis and web search analysis, we propose the “high-risk social contact index (HR-SCI)” metric, which takes into consideration both the density of population and risk of COVID-19 (Fig. 1), in contrast to previous metrics that only rely on either mobility data or web search data alone. The methods are tested using data of >450,000 users in Tokyo, who were observed across a 7-month period (February 1, 2020 to September 10, 2020). Experiments show that the HR-SCI is capable of predicting COVID-19 outbreak hotspots 1–2 weeks before the official observations of the outbreak. The HR-SCI metric can be used to develop early warning systems for COVID-19 hotspots that inform government agencies the locations of potential outbreaks, allowing them to plan effective prevention and preparedness strategies.

The key contributions of this paper are as follows:

- This study is the first to test the usage of web search and mobility data, which are linked by user IDs, for the prediction of COVID-19 outbreak hotspots.
- We propose a novel metric, high risk social contact index (HR-SCI), that captures both the social contact density and the COVID-19 contraction risk levels of the users, with high spatio-temporal granularity.
- We verify that the HR-SCI can improve the predictability of COVID-19 hotspot locations through several case studies, compared to using just the social contact index or web search queries alone.

2. Related works

2.1. Mobility analysis during COVID-19

Mobile devices have become ubiquitous in many areas around the world, providing opportunity to analyze human mobility dynamics in an unprecedented spatial and temporal granularity and scale (Blondel, Decuyper, & Krings, 2015). To monitor and evaluate the impacts of such interventions, large scale mobile phone data has been utilized as an effective data source (Oliver et al., 2020). Studies on human mobility analysis have used such data to model disease dynamics (Bengtsson et al., 2015; Finger et al., 2015; Tizzoni et al., 2014; Wesolowski et al., 2012). During the COVID-19 crisis, researchers, industry, and government agencies have utilized large-scale mobility datasets to evaluate the effects of NPIs in various countries, including the United States (Gao et al., 2020; Klein et al., 2020; Wellenius et al., 2020), the United Kingdom (Santana et al., 2020), Italy (Bonaccorsi et al., 2020; Bonato et al., 2020; Cintia et al., 2020; Pepe et al., 2020), China (Kraemer et al., 2020; Lai et al., 2020), Sweden (Dahlgberg et al., 2020), Germany (Schlosser et al., 2020), Spain (Orro, Novales, Montaegudo, Pérez-López, & Bugarn, 2020), Austria (Heiler et al., 2020), and Japan (Mizuno & Ohishi, 2020; Yabe et al., 2020). None of these studies have attempted to integrate the analysis of web search queries to predict COVID-19 outbreak hotspots.

2.2. Applications of web search data analysis

Mining of web search query data has attracted the attention of researchers and practitioners ever since search engines were introduced to the world (Silverstein, Marais, Henzinger, & Moricz, 1999). Web search data has been utilized for various applications, for example, to predict users’ demographics (Wu et al., 2019) and mobility decisions during crisis (Yabe, Tsoubouchi, Shimizu, Sekimoto, & Ukkusuri, 2019). During the COVID-19 pandemic, many studies have utilized web search query data to understand information seeking behavior and the occurrence of infodemics (Bento et al., 2020; Mavragani, 2020; Rovetta & Bhagavathula, 2020). Others have used such data to detect the increase in COVID-19 symptoms (e.g. loss of smell) (Rajan et al., 2020; Walker, Hopkins, & Surda, 2020), and also for predicting outbreaks (Hisada et al., 2020; Li et al., 2020). With the availability of open datasets (Xu et al., 2020), there is great potential in further using web search query data for pandemic response and prevention. Despite the vast array of studies, none have attempted to integrate web search data with mobility data to predict COVID-19 outbreaks.

3. Context: COVID-19 in Tokyo, Japan

Japan has experienced a low number of cases and deaths due to COVID-19 in comparison to other countries in Europe and America, despite the social and physical proximity to China and intervention policies that are not as aggressive as some of the other countries (Dong, Du, & Gardner, 2020). Non-pharmaceutical interventions implemented by the Japanese government include a non-mandatory closure and remote-working of non-essential business employees (February 26th), closures of public elementary, junior high and high schools (March 2nd), and inbound entry restrictions, starting with Hubei Province, China (February 3rd), until restricting inbound visitors from 73 countries (April 3rd). No mandatory lockdowns were enforced in Japan. The State of Emergency (SoE) was declared on April 7th, and was lifted on May 25th, after observing a significant decrease in cases (see Fig. 2, gray bars). Although Japan was able to contain the disease, many cities started to see an increase of cases in early July, and experienced the second wave in July to September (see Fig. 2, gray bars). Tokyo, which has the largest number of COVID-19 cases among prefectures in Japan, had around 400 new daily cases at its peak. As of October 17th, Tokyo has had 28,839 cases (out of 90,979 in Japan) and 434 deaths (out of 1650 in Japan).

4. Data

In this study, we utilized web search data and GPS location data, that are linked with the same user identifiers (IDs), owned by Yahoo Japan Corporation. Because the user IDs are linked, we are able to 1) identify users who have a high risk of COVID-19 contraction from their web search behavior, and 2) track their mobility patterns to measure their social contact rates with other users in the city.

4.1. Privacy policies for user data

Yahoo Japan Corporation (YJ) has developed its own privacy policy and requires users to read and agree to its privacy policy before using any of the services provided by YJ. Furthermore, because location data is highly sensitive for the users, users were asked to sign an additional consent form specific to the collection and usage of location data when they used apps that collect location information. The additional consent

https://about.yahoo.co.jp/info/en/
explains the frequency and accuracy of location information collection, and also the purpose and how the data will be used. In addition to the above consent, YJ asked for additional consent from the users in this study because the analysis related to COVID-19 and personal health are much more sensitive. Therefore, YJ performed a double consent process, where the users who have given their consent to the usage of location information and web search queries were asked again, if they wish to provide their consent to be included in the dataset.

Moreover, YJ implemented strict restrictions in the analysis procedure. The methodology for handling the data and for obtaining user consent for this study were supervised by an advisory board composed of external experts. YJ also ensured that research institutions other than YJ that participate in this study do not have direct access to the data. Although external research institutions were allowed to analyze aggregated data, the actual raw data were kept within YJ, and any analysis performed on raw data were performed within servers administered by YJ. In summary, given the high sensitivity of the study, this study was performed with careful consideration of the users’ privacy.

4.2. Individual human mobility data

GPS location data are anonymized so that individuals cannot be specified, and personal information such as gender, age and occupation are unknown. Each GPS location record contains the user’s unique ID, timestamp of the observation, longitude, and latitude. The data has a sample rate of approximately 2% of the entire population. The data acquisition frequency of GPS locations varies according to the movement speed of the user to minimize the burden on the user’s smartphone battery. If it is determined that the user is staying in a certain place for a long time, data is acquired at a relatively low frequency, and if it is determined that the user is moving, the data is acquired more frequently. We overcome this varying data acquisition frequency by spatially and temporally interpolating the location data, as explained in Section 5.3. A panel of users who were active each day in Tokyo metropolitan area before, during and after the COVID-19 pandemic were selected from the pool of users that have agreed to the aforementioned consents. This led to a sample of about 450,000 users, with approximately 50 observation points per user each day.

4.3. Individual web search data

In addition to GPS location data, Yahoo Japan collects the web search queries of the users to improve the web search quality. Each web search query record contains the user’s unique ID, timestamp of the search, and query text that was searched. The web search data were used to identify users who have a high risk of COVID-19 contraction, following the methods explained in Section 5.2.

4.4. Number of COVID-19 cases in Tokyo

The daily number of confirmed COVID-19 cases in Tokyo were reported by the Tokyo metropolitan government through their Github data portal (Tokyo Metropolitan Government, 2020) (gray bars in Fig. 2).
users who intensively search about COVID-19 symptoms (Section 5.2).

5.1.2. Definition 2 (high risk user)

A high risk user is defined as a user who conducts more web search sessions that are related to COVID-19 symptoms than the pre-defined threshold.

5.1.3. Definition 3 (social contact index)

The social contact index (SCI) measures the average amount of encounters that each user experiences due to movements outside of their estimated home locations. When \(n\) individuals are observed in the given spatial unit and temporal window, the number of contacts in this cell can be quantified as 

\[
C_2 = \frac{n!}{2!(n-2)!}\]

5.1.4. Definition 4 (high risk social contact index)

The high risk social contact index (HR-SCI) is a composite measure of the SCI and the number of high risk users. HR-SCI measures the average amount of social contacts that high risk users encounter due to movements outside of their estimated home locations.

5.2. Scoring of web search queries

To identify high risk users of COVID-19 contraction, the users’ web search queries from Yahoo Japan Search were given risk scores. Queries submitted by users were treated as “Covid-19 queries” if they match against pre-defined query patterns. The query patterns consist of 3 types of query phrases: 1) COVID-19 symptom related queries, 2) names of medical institutions related to COVID-19 care, and 3) names of locations, as shown in Table 1. For the first group of queries, 186 COVID-19 symptom related queries were determined, such as ‘coronavirus high fever’, ‘no smell’, ‘may have corona’ including various slang words (e.g., ‘corona’ instead of ‘coronavirus’, which is typically used by Japanese-speaking users). For the second group of queries, 5 queries that represent medical institutions concerned with COVID-19 were determined. These include facilities such as hospitals designated by the local health authorities to be specialized for the treatment of COVID-19 infected patients. The third group of queries include 2168 queries representing names of locations (e.g., Shibuya). The latter two groups of queries only were scored as “COVID-19 related” only when they were searched together to qualify as a “COVID-19 related search”.

Table 1

| Type                  | Quantity | Examples                      |
|-----------------------|----------|-------------------------------|
| Symptom related       | 186      |“coronavirus high fever”, “no smell”, “may have corona” |
| Medical institutions  | 5        | “hospital”, “medical clinic”  |
| Location names        | 2168     | “Shibuya”, “Shinjuku”         |

*queries on medical institution names and location names need to be searched together to qualify as a “COVID-19 related search”.

Fig. 2. Time series of estimated COVID-19 related metrics aggregated at the entire Tokyo metropolitan scale between February 1st and September 5th, 2020 (left axis). A) Total number of high risk users. B) Social contact index of an average user. C) High risk social contact index, which is the total amount of social contacts the high risk users encounter. Gray bars in each panel show the daily number of new COVID-19 cases in Tokyo (right axis).
risk users with \( k = 3 \) in the Tokyo region, along with the daily number of new cases.

5.3. Social contact analysis

Using the GPS location data of each user, we compute the social contact index within a given region on the time of interest. To overcome data sparsity caused by battery saving functions of the app (explained in Section 4.2), we perform spatio-temporal interpolation of the GPS location observations. Because the GPS data are collected less frequently when movement is detected, we assume that the individual users are static while there are no observations. In the final interpolated dataset, each individual’s locations are recorded every 30 min. Using the interpolated individual trajectory data produced from mobile phone location data, the social contact indexes were computed. The social contact indexes shown in Fig. 2B were computed for 30 min intervals. First, for polulated individual trajectory data produced from mobile phone location static while there are no observations. In the final interpolated dataset, when movement is detected, we assume that the individual users are data sparsity caused by battery saving functions of the app (explained in Section 4.2), we perform spatio-temporal interpolation of the GPS contact index within a given region on the time of interest. To overcome new cases.

Fig. 2B shows the SCI between February 1st and September 5th in the Tokyo region, along with the daily number of COVID-19 cases. The results indicate that the two metrics – high risk users and high risk social contact index – have similar high predictability of the daily number of cases when inspecting the two waves separately, however, the High Risk SCI outperforms the HRU index when we perform lagged correlation analysis on both waves. The time lagged time series data are visualized in Fig. 3D-F, where the dotted black line shows the original observation, solid black line shows the lagged waves in aggregate (“both waves”), and colored dashed and solid lines show the time lagged first wave and second wave predictions, respectively. The peak cross correlation are shown in Table 2, and indicates that while HRU and HR-SCI metric perform better, SCI itself is not able to capture the trends of the number of new cases. Although the HRU metric performed the best during the first wave with \( R = 0.894 \) and lag of 8 days, the HR-SCI performed the best during the second wave with \( R = 0.756 \) and for the two waves combined with \( R = 0.641 \). From these results, we conclude that the HR-SCI metric perform well to predict the temporal trends of new daily COVID-19 cases across multiple waves.

6.2. Microscopic outbreak hotspot prediction

In the previous section, it was shown that the HR-SCI metric was most effective in predicting the macroscopic (Tokyo metropolitan area scale) prediction of COVID-19 outbreak across multiple waves. Policy makers could further benefit from more microscopic, finer spatially-grained prediction (and early warning system) of outbreak hotspots. In this section, we tested the effectiveness of the three metrics on predicting outbreak hotspots beforehand, in the microscopic (i.e., 125 m grid) spatial scale.

First, we visualize the three indexes (SCI, HRU, and HR-SCI) for each km × 1 km grid in Tokyo metropolitan region. Fig. 4 shows the spatial distributions of the three index scores (SCI, HRU, HR-SCI) on the 1 km × 1 km grid scale, aggregated between June 1st and July 31st, and normalized between 0 and 1. From Table 3, we observe high SCI scores in major central business districts including Shinjuku, Tokyo, and Ikebukuro, whereas the high risk index is more dispersed, with lower density areas such as Hibiya, Hamamatsu-cho, and Ogikubo ranked higher compared to SCI. HR-SCI is a composite measure, including locations from both the SCI and HRU ranking. As we can see from the normalized scores of the top 100 ranked locations of each metric in Fig. 5, HR-SCI is more selective, showing high scores on only a few grid cells.

To validate the effectiveness of the HR-SCI scores on predicting outbreak hotspot locations, we collected news articles reporting the occurrence of COVID-19 outbreaks in Tokyo. We found news reports of outbreak events in eight areas in Tokyo – Shinagawa, Mizonokuchi, Ikebukuro, Nishikasai, Tachikawa, Oshimachi, Kinsicho, and Kitasenju – which were used to validate the predictability of outbreak hotspots using the metrics. The SCI and HR-SCI values of the grid cell that contains the main location of each area were used. For example for Shinagawa, the 125 m × 125 m cell that contains the central business district around Shinagawa Station was used.
Computers, Environment and Urban Systems 92 (2022) 101747

6.2.1. Shinagawa area

Shinagawa (Fig. 6A) is known as one of the largest central business districts in Tokyo, and therefore consistently has a high social contact index, ranking between 8th to 10th in all periods. However, we can also observe that Shinagawa ranks very low (between 100th to 1000th) using the high risk user count index. On the other hand, Shinagawa’s high risk social contact index rank fluctuates over time, and ranks 5th (week of July 6th) and 7th (week of July 13th). It is reported that cases in Shinagawa increased from July 13th to August 2nd, and has had several COVID-19 clusters during that period. Therefore, using SCI would result in false positive predictions until mid July, whereas HR-SCI effectively predicts outbreaks 1–2 weeks beforehand.

6.2.2. Ikebukuro area

Similar to the Shinagawa area, Ikebukuro is also a bustling area with many shopping and business facilities, and ranks constantly in the top 10 using SCI (between 2nd and 7th in entire period) (Fig. 6B). However, Ikebukuro never ranks in the top 50 using high risk users index. Using HR-SCI, Ikebukuro ranks in the top 10 twice during the study period: the weeks of June 22nd and July 20th (Fig. 6C). Indeed, COVID-19 outbreaks have been reported on July 4th and July 30th in the area, which are both 2 weeks after the area was ranked 10th using the HR-SCI. Ikebukuro showcases another example where using SCI, it would constantly be false-positively predicted as outbreak locations, but using HR-SCI, outbreak timings can be accurately predicted 1–2 weeks beforehand.

6.2.3. Mizonokuchi area

Mizonokuchi area is a residential area located in the suburbs of Tokyo, unlike Shinagawa and Ikebukuro. As shown in Fig. 6C, Mizonokuchi never ranks in the top 10 using social contact index, due to low population density. However, we can also observe that Shinagawa ranks very low (between 100th to 1000th) using the high risk user count index. On the other hand, Shinagawa’s high risk social contact index rank fluctuates over time, and ranks 5th (week of July 6th) and 7th (week of July 13th). It is reported that cases in Mizonokuchi increased from July 13th to August 2nd, and has had several COVID-19 clusters during that period. Therefore, using SCI would result in false positive predictions until mid July, whereas HR-SCI effectively predicts outbreaks with 2 weeks in hand.

6.2.4. Nishi-Kasai area

Nishi-Kasai area, similar to Mizonokuchi area, is a residential area with smaller active population compared to Shinagawa and Ikebukuro. Therefore, the area always ranks below 20th in SCI. However, using the HR-SCI, Nishi-Kasai ranks in the top 10 twice during the study period: the weeks of June 22nd and July 20th (Fig. 6D). Indeed, COVID-19 outbreaks have been reported on July 4th and July 30th in the area, which are both 2 weeks after the area was ranked 10th using the HR-SCI. Nishi-Kasai is another case where because of the low active population density, the area never appears in the top rankings using SCI, but and again ranks in the top 10 during the end of July (weeks of July 20th and 27th). It has been reported that in Ikebukuro, the number of cases have risen from mid-June to the end of June, and by the end of June had the most number of new cases in Tokyo. Moreover, occurrence of COVID-19 clusters in large shopping malls and the city hall have been reported at around early August (which is 1–2 weeks after the second increase in HR-SCI). Ikebukuro showcases another example where using SCI, it would constantly be false-positively predicted as outbreak locations, but using HR-SCI, outbreak timings can be accurately predicted 1–2 weeks beforehand.

Table 2

| Metric                  | Pearson Correlation (lagged days) |
|------------------------|-----------------------------------|
| High Risk Users SCI    | 0.894 (±8 days)                   |
| High Risk SCI          | 0.641 (±16 days)                  |
| Both waves combined    | 0.616 (±9 days)                   |

Fig. 3. Time lagged cross correlation analysis of the high risk users, social contact index, and high risk social contact index metrics against the daily number of new cases. Results indicate higher predictability using HRU and HR-SCI; the metrics preceded the daily cases trend by 8–16 days during the first wave, and by 16–23 days during the second wave. HR-SCI had a better overall fit on the two waves.

6.2.1. Shinagawa area

Shinagawa (Fig. 6A) is known as one of the largest central business districts in Tokyo, and therefore consistently has a high social contact index, ranking between 8th to 10th in all periods. However, we can also observe that Shinagawa ranks very low (between 100th to 1000th) using the high risk user count index. On the other hand, Shinagawa’s high risk social contact index rank fluctuates over time, and ranks 5th (week of July 6th) and 7th (week of July 13th). It is reported that cases in Shinagawa increased from July 13th to August 2nd, and has had several COVID-19 clusters during that period. Therefore, using SCI would result in false positive predictions until mid July, whereas HR-SCI effectively predicts outbreaks with 2 weeks in hand.

6.2.2. Ikebukuro area

Similar to the Shinagawa area, Ikebukuro is also a bustling area with many shopping and business facilities, and ranks constantly in the top 10 using SCI (between 2nd and 7th in entire period) (Fig. 6B). However, Ikebukuro never ranks in the top 50 using high risk users index. Using HR-SCI, Ikebukuro ranks 2nd, 1st, 1st in the three weeks from June 1st, and again ranks in the top 10 during the end of July (weeks of July 20th and 27th). It has been reported that in Ikebukuro, the number of cases have risen from mid-June to the end of June, and by the end of June had the most number of new cases in Tokyo. Moreover, occurrence of COVID-19 clusters in large shopping malls and the city hall have been reported at around early August (which is 1–2 weeks after the second increase in HR-SCI). Ikebukuro showcases another example where using SCI, it would constantly be false-positively predicted as outbreak locations, but using HR-SCI, outbreak timings can be accurately predicted 1–2 weeks beforehand.

6.2.3. Mizonokuchi area

Mizonokuchi area is a residential area located in the suburbs of Tokyo, unlike Shinagawa and Ikebukuro. As shown in Fig. 6C, Mizonokuchi never ranks in the top 10 using social contact index, due to low population density. However, we can also observe that Shinagawa ranks very low (between 100th to 1000th) using the high risk user count index. On the other hand, Shinagawa’s high risk social contact index rank fluctuates over time, and ranks 5th (week of July 6th) and 7th (week of July 13th). It is reported that cases in Shinagawa increased from July 13th to August 2nd, and has had several COVID-19 clusters during that period. Therefore, using SCI would result in false positive predictions until mid July, whereas HR-SCI effectively predicts outbreaks with 2 weeks in hand.

6.2.4. Nishi-Kasai area

Nishi-Kasai area, similar to Mizonokuchi area, is a residential area with smaller active population compared to Shinagawa and Ikebukuro. Therefore, the area always ranks below 20th in SCI. However, using the HR-SCI, Nishi-Kasai ranks in the top 10 twice during the study period: the weeks of June 22nd and July 20th (Fig. 6D). Indeed, COVID-19 outbreaks have been reported on July 4th and July 30th in the area, which are both 2 weeks after the area was ranked 10th using the HR-SCI. Nishi-Kasai is another case where because of the low active population density, the area never appears in the top rankings using SCI, but
6.3. Sensitivity analysis

To understand the effects of the spatial and temporal threshold parameters that we selected to compute the social contacts (125 m and 30 min time intervals) in the previous experiments, we conducted a parameter sensitivity check for the spatial threshold and temporal thresholds. Fig. 7 shows the sensitivity of the high risk social contact index when computed using different temporal (30 min, 1 h, 2 h) and spatial (1 km, 500 m, 250 m, 125 m) threshold parameters. Note that since the original data contained the location information of each individual every 30 min, the minimum temporal threshold was 30 min.

The overall patterns computed using all spatial and temporal threshold parameters present similar trends, where we see two significant peaks; 2020 April just before the first wave, and 2020 July, just before the second wave. Changing the temporal threshold between 30 min, 60 min, and 120 min changes the granularity of the estimations, where we see a more smoothed out time series data when social contacts are computed in 120 min thresholds, which is intuitive. When changing the spatial threshold parameter, we observe interesting patterns especially in the first wave. The high risk SCI is more sensitive to the spatial threshold parameter during the first wave, where reducing the contact threshold from 1 km to 125 m reduces the HR social contact index by around 70%. However, the second wave is significantly less affected by the change in the spatial threshold parameter. This indicates that during the first wave, high risk individuals had higher spatial proximity with each other,

accurately appears before actual outbreaks using HR-SCI.

**Table 3**
Top 10 locations in Tokyo metropolitan area with highest SCI, HRU, and HR-SCI indexes.

| Rank | SCI        | High Risk Users | High Risk SCI |
|------|------------|-----------------|---------------|
| 1    | West Shinjuku | Hibiya          | Tokyo         |
| 2    | Tokyo      | Tokyo           | West Shinjuku |
| 3    | Ikebukuro  | Ikebukuro       | Akihabara     |
| 4    | Akihabara  | Akihabara       | Hibiya        |
| 5    | Ginza      | Hamamatsu-cho   | Ikebukuro     |
| 6    | Ohtemachi  | Shinbashi       | Kitasenju     |
| 7    | Hibiya     | Shimokitazawa   | Ohtemachi     |
| 8    | Kitasenju  | West Shinjuku   | Shinbashi     |
| 9    | South Shibuya | Ogikubo     | Ogikubo       |
| 10   | Shinagawa  | Ohtemachi       | Hamamatsu-cho |

**Fig. 4.** Spatial plot of the total (A) social contact index, (B) high risk user count, and (C) high risk social contact index during June 1st - July 31st in the metropolitan Tokyo area, aggregated into 1 km grid cells.

**Fig. 5.** Normalized scores of each metric in 1 km grid cells sorted by their rank, showing high skewness of the HR-SCI.
where many of them could be reached within 1 km from each other. On the other hand, the high risk individuals were more dispersed and distanced during the second wave, posing less contact risks.

To further understand the effects of the spatial and temporal threshold parameters on the early warning performance of the 2 COVID-19 waves, the lagged correlation analysis was conducted for each of the 12 (= 4 spatial \(\times\) 3 temporal) threshold parameter pairs. Fig. 8 plots the lagged correlation coefficient (top row) and the time lag (bottom row).

Fig. 9 shows the lagged correlation analysis plots for the parameter combinations. The time lag for the 1st wave, 2nd wave, and both waves between the high risk social contact index and the number of cases are shown. We observe that the correlation is higher when we compute the index using more spatially granular information. Interestingly, the right plot on the top row shows that for both waves, the temporal threshold of 60 min performs best (also for the second wave). The time lags under different parameters are mostly stable for each wave.
Computers, Environment and Urban Systems 92 (2022) 101747

7. Discussion

7.1. Advantages and disadvantages of proposed metrics

In this study, we investigated the effectiveness of various metrics for predicting outbreak hotspot locations of COVID-19, using a unique dataset that contains both the web search queries and GPS location information of users. The two experiment setups – predicting the macroscopic (Tokyo metropolitan area) and microscopic (outbreaks in 125 m cell scale) – unraveled the strengths and weaknesses of the three metrics: high risk user (HRU) count, social contact index (SCI), and high risk social contact index (HR-SCI).

It was shown in Fig. 3 and Table 2 that HRU is fairly effective in capturing the macroscopic trends of COVID-19 outbreaks for each of the waves independently. We observed high correlation between the daily number of mobile phone users who had high number of COVID-19 related web search sessions and the daily new cases count provided by the Tokyo Metropolitan Government. In particular, cross correlation analysis showed that the fluctuations in the high risk user count preceded the cases count by 1–2 weeks, which can be useful in predicting the outbreaks in the macroscopic scale. However, the HRU index was not effective in predicting outbreak locations in the microscopic scale, as shown in Section 6.2. This is because on the microscopic scale (125 m grid cells), the decrease in the average number of users observed in each cell increases the noise in the data. These metrics could become biased even by small groups of people who search about COVID-19 out of pure interest.

The social contact index (SCI), which is used in many studies, could be effective in predicting outbreaks if COVID-19 cases occur in high dense areas, since essentially SCI captures the population density of that area. However, as shown in the analysis in Section 6.2, SCI has two major drawbacks in predicting hotspots. First, the SCI is relatively static over time, staying consistently large in locations that are more congested, including central business districts and hub stations (e.g., Tokyo, Shibuya, Shinjuku), and consistently low in more rural areas. Although it is true that COVID-19 has a higher probability of spreading in high-density areas, because of this static nature of the SCI metric, it is difficult to predict the timings of such outbreaks. In the case of Shinagawa (one of the busiest central business districts in Tokyo), outbreaks could not be detected using SCI because the area was always congested throughout the COVID-19 crisis. Mizonokuchi (a town in the outskirts of Tokyo) which had an outbreak also could not be detected using SCI because of the consistently low SCI.

Considering the pros and cons of the HRU and SCI metrics, the high risk SCI (HR-SCI) metric is able to capture both 1) existence of high risk users and 2) high population (contact) density. Thus, especially in the microscopic scale, HR-SCI was shown to be more effective in predicting hotspot locations in Section 6.2. Therefore, HR-SCI could be utilized to implement early warning systems for COVID-19 outbreak hotspot locations.

7.2. Limitations and future works

The presented empirical results should be considered in the light of some limitations. First, we found that although the HRU and HR-SCI metrics had high predictability of COVID-19 cases in Fig. 3, the optimal time lags in the two waves varied (first wave: 9–16 days, second wave: 16–23 days). Many reasons could be causing this difference, however, this suggests a change in the the web search behavior of the users between the two waves. As the virus spreads and more information becomes available due to the increase of cases, the type of symptoms that users search in relation to COVID-19 could increase (e.g., it was unknown that loss of smell was related to COVID-19 in the first weeks of the outbreak). Moreover, the difference in the absolute amount of high risk users between the two waves show that awareness levels of COVID-19 had also changed, and that as people become more used to the
Fig. 9. Sensitivity of the lagged correlation between the high risk social contact index and number of cases for different temporal (30 min, 1 h, 2 h) and spatial (1 km, 500 m, 250 m, 125 m) threshold parameters.
disease, less people search about COVID-19 symptoms. Despite these issues, the finding that we can predict the outbreak around 1–2 weeks beforehand is valuable and has various potential applications as discussed later. Future research could focus more on the modeling of the web search behavior to improve the accuracy of the outbreaks.

Second, computational costs increase drastically as we increase the spatial resolution of the analysis. For example, in Tokyo, there are around 2400 1 km × 1 km grid cells, which were further divided into over 150,000,125 m × 125 m grid cells. Running the analysis with high computing resources could be one solution to this issue. Finally, collecting ground truth data for COVID-19 outbreaks on the microscopic spatial scales were challenging, with information sources scattered across websites of various institutions and agencies. Constructing a comprehensive database that contains when, where, and how many patients were reported on fine spatial scales (with careful consideration of different regions around the world) could be an interesting extension to this study.

3. Potential applications

The usability of high risk social contact index is not limited to COVID-19, and can be applied to any contagious disease in theory. For example, flu has been the subject of many previous studies on predicting disease spread using web search data and mobility data (e.g., (Lampos & Cristianini, 2010; Fanigutti, Tizzoni, Bajardi, Smoreda, & Colizza, 2017)). Although the web search intensity, frequency, and the disease-related parameters (e.g., effective reproduction number) could be different in other diseases, it would be valuable to test this method on other epidemics. Moreover, testing the applicability of this method in different regions around the world could be an interesting extension to this study.

In addition to the early warning system mentioned previously, these metrics can be used to develop a personalized alert and navigation app on smartphones. Although there are existing apps that inform users about their risk of contracting the disease based on their mobility information (e.g., COCOA app developed by the Japanese Government), they do not provide early warning information on where potential outbreaks could occur. Using the high risk social contact index, we would be able to provide navigation and personalized early warning to users so that they can avoid visiting or passing through high risk areas.

8. Conclusion

As COVID-19 continues to affect public health in cities across the world, early warning systems that can predict where the next outbreak would occur is of significant importance for government agencies. In this study, we used web search data and GPS location data, which are linked with common user IDs, to predict outbreak hotspot locations using the high risk social contact index. Validation using data from Tokyo, Japan showed that compared to previously proposed metrics, the high risk social contact index is capable of predicting the timing of outbreaks 1–2 weeks beforehand in a microscopic (125 m) spatial scale. This study proposes a novel method to predict disease outbreak hotspots, which can be used to develop early warning systems that may assist government agencies to prepare effective strategies for disease spread prevention.

Acknowledgements

T.Y. and S.V.U. were partly funded by NSF Grant No. 1638311 CRISP Type 2/Collaborative Research: Critical Transitions in the Resilience and Recovery of Interdependent Social and Physical Networks.

References

https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/cocoa_00138.html

7 https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/cocoa_00138.html

Bengtsson, L., Gaudart, J., Lu, X., Moore, S., Wetter, E., Salathé, R., & Piarrouts, R. (2015). Using mobile phone data to predict the spatial spread of cholera. Scientific Reports, 5, 8925.

Bento, A. L., Nguyen, T., Wing, C., Lozano-Rojas, F., Ahn, Y. Y., & Simon, K. (2020). Evidence from internet search data shows information-seeking responses to news of local COVID-19 cases. Proceedings of the National Academy of Sciences, 117, 111220–111222.

Blondel, V. D., Decuyper, A., & Kring, K. (2015). A survey of results on mobile phone datasets analysis. EFO data science, 4, 16.

Bonacorsi, G., Piéri, F., Cinelli, M., Flori, A., Galeazzi, A., Porcelli, F., Schmidt, A. L., Valense, C. M., Scala, A., Quattrociocchi, W., et al. (2020). Economic and social consequences of human mobility restrictions under covid-19. Proceedings of the National Academy of Sciences, 117, 15530–15535.

Bonato, P., Gintis, F., Fabbrini, F., Fadda, D., Giannotti, F., Lopalo, P. I., … Pedreschi, D., et al. (2020). Mobile phone data analytics against the covid-19 epidemics in Italy: Flow diversity and local job markets during the national lockdown. arXiv preprint arXiv:2004.11278.

Cintia, P., Fadda, D., Giannotti, F., Pappalardo, L., Rossetti, G., Pedreschi, D., … Peone, V., et al. (2020). The relationship between human mobility and viral transmissibility during the covid-19 epidemics in Italy. arXiv preprint arXiv:2006.03141.

Dahlberg, E., Eden, P. A., Grenyer, E., Lyhagen, J., Östh, J., Siretky, A., & Toger, M. (2020). Effects of the covid-19 pandemic on population mobility under mild policies: Causal evidence from Sweden. arXiv preprint arXiv:2004.09087.

Dodge, S., Su, R., Johnson, J., Simharoon, A., Goulais, K., Smith, J. L., & Ahearn, S. C. (2021). Ortega: An object-oriented time-geographic analytical approach to trace space-time contact patterns in movement data. Computers, Environment and Urban Systems, 88, Article 101630.

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. The Lancet infectious diseases, 20(5), 523–524.

Fan, J., & Stewart, K. (2021). Understanding collective human movement dynamics during large-scale events using big geosocial data analytics. Computers, Environment and Urban Systems, 87, Article 101605.

Finger, F., Genolet, T., Mars, L., de Magny, G. C., Manga, N. M., Rinaldo, A., … Bertuzzo, E. (2016). Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks. Proceedings of the National Academy of Sciences, 113, 6421–6426.

Flaxman, S., Mishra, S., Gandy, A., Urwin, H. J. T., Coupland, H., Melan, T. A., Guzman, P. N., et al. (2020). Estimating the number of infections and the impact of non-pharmaceutical interventions on covid-19 in europe: Technical description update. arXiv preprint arXiv:2004.11342.

Gao, S., Rao, J., Kang, Y., Liang, Y., & Kruse, J. (2020). Tracking the flu pandemic by monitoring the social web. In Proceedings of the National Academy of Sciences, 113, 533–534.

Guzman, P. N., et al. (2020). Assessing changes in commuting and individual mobility from smartphones against covid-19 clusters in Japan. arXiv preprint arXiv:2004.10100.

Hanage, W. P., et al. (2020). Assessing changes in commuting and individual mobility during large-scale events using big geosocial data analytics. Computers, Environment and Urban Systems, 87, Article 101605.

Hillier, F., Genset, T., Mars, L., de Magny, G. C., Manga, N. M., Rinaldo, A., … Bertuzzo, E. (2015). Social contact index is capable of predicting the timing of outbreaks 1–2 weeks beforehand in a microscopic (125 m) spatial scale. This study proposes a novel method to predict disease outbreak hotspots, which can be used to develop early warning systems that may assist government agencies to prepare effective strategies for disease spread prevention.

Acknowledgements

T.Y. and S.V.U. were partly funded by NSF Grant No. 1638311 CRISP Type 2/Collaborative Research: Critical Transitions in the Resilience and Recovery of Interdependent Social and Physical Networks.
Orro, A., Novales, M., Monteagudo, Á., Fèrre-López, J. B., & Bugarn, M. R. (2020). Impact on city bus transit services of the covid–19 lockdown and return to the new normal: The case of a coruna (Spain). Sustainability, 12, 7206.

Panigutti, C., Tizzoni, M., Bajardi, P., Smoreda, Z., & Colizza, V. (2017). Assessing the use of mobile phone data to describe recurrent mobility patterns in spatial epidemic models. Royal Society Open Science, 4, Article 160955.

Pepe, E., Bajardi, P., Gauvin, L., Privitera, F., Lake, B., Cattuto, C., & Tizzoni, M. (2020). Covid-19 outbreak response: A first assessment of mobility changes in Italy following national lockdown. medRxiv. https://doi.org/10.1101/2020.03.22.20039933

Rajan, A., Sharaf, R., Brown, R. S., Sharaiha, R. Z., Lebwohl, B., & Mahadev, S. (2020). Association of search query interest in gastrointestinal symptoms with covid-19 diagnosis in the United States: Infodemiology study. JMIR Public Health and Surveillance, 6, Article e19354.

Rovetta, A., & Bhagavathula, A. S. (2020). Covid-19-related web search behaviors and infodemic attitudes in Italy: Infodemiological study. JMIR Public Health and Surveillance, 6, Article e19374.

Santana, C., Botta, F., Barbosa, H., Privitera, F., Menezes, R., & Di Clemente, R. (2020). Analysis of human mobility in the UK during the covid-19 pandemic. https://covid19-uk-mobility.github.io/.

Schlosser, F., Maier, B. F., Hinrichs, D., Zachariae, A., & Brockmann, D. (2020). Covid-19 lockdown induces structural changes in mobility networks – Implication for mitigating disease dynamics. arXiv preprint arXiv:2007.01583.

Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999). Analysis of a very large web search engine query log. In ACM SIGIR Forum (pp. 6–12). NY, USA: ACM New York.

Tizzoni, M., Bajardi, P., Decuypere, A., King, G. K. K., Schneider, C. M., Blondel, V., & Colizza, V. (2014). On the use of human mobility proxies for modeling epidemics. PLoS Computational Biology, 10.

Tokyo Metropolitan Government. (2020). Covid-19 information website. https://github.com/tokyo-metropolitan-gov/covid19 (accessed October 16, 2020).

Ukkusuri, S. V., Yabe, T., & Seetharam, K. (2020). Non-pharmaceutical interventions for covid-19: Evidence from large-scale mobility data in Tokyo. ADBI Policy Brief 2020–6.

Walker, A., Hopkins, C., & Surda, P. (2020). The use of google trends to investigate the loss of smell related searches during covid-19 outbreak. In International Forum of Allergy & Rhinology. Wiley Online Library.

Wellensius, G. A., Vispute, S., Espinosa, V., Fabrikanth, A., Tsai, T. C., Hennessey, J., … Pearce, A., et al. (2020). Impacts of state-level policies on social distancing in the United States using aggregated mobility data during the covid-19 pandemic. arXiv preprint arXiv:2004.10172.

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. Science, 338, 267–270.

Wu, C., Wu, F., Liu, J., He, S., Huang, Y., & Xie, X. (2019). Neural demographic prediction using search query. In Proceedings of the twelfth ACM international conference on web search and data mining (pp. 654–665).

Xu, B., Gutierrez, B., Mekara, S., Sewalk, K., Goodwin, L., Loskill, A., Cohn, E. L., Hosven, Y., Hill, S. C., Cobo, M. M., et al. (2020). Epidemiological data from the covid-19 outbreak, real-time case information. Scientific Data, 7, 1–6.

Yabe, T., Tsuiboichi, K., Fujisawa, N., Wada, T., Sekimoto, Y., & Ukkusuri, S. V. (2020). Non-compulsory measures sufficiently reduced human mobility in Tokyo during the covid-19 epidemic. Scientific Reports, 10, 1–9.

Yabe, T., Tsuiboichi, K., Shimizu, T., Sekimoto, Y., & Ukkusuri, S. V. (2019). Predicting evacuation decisions using representations of individuals’ pre-disaster web search behavior. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2707–2717).

Zhang, J., Litvinova, M., Liang, Y., Wang, Y., Wang, W., Zhao, S., … Vespignani, A., et al. (2020). Changes in contact patterns shape the dynamics of the covid-19 outbreak in China. Science.