Using mobile network data to color epidemic risk maps

Elisa Cabana
elisa.cabana@imdea.org
IMDEA Networks Institute
Madrid, Spain

Andra Lutu
andra.lutu@telefonica.com
Telefonica Research
Barcelona, Spain

Enrique Frias-Martinez
enrique.frias@ucjc.edu
Universidad Camilo Jose Cela
Madrid, Spain

Nikolaos Laoutaris
nikolaos.laoutaris@imdea.org
IMDEA Networks Institute
Madrid, Spain

ABSTRACT

In this paper we propose a method for using mobile network data to detect potential COVID-19 hospitalizations and derive corresponding epidemic risk maps. We apply our methods to a dataset from more than 2 million cellphones, collected over the months of March and April in 2020 by a British mobile network provider. The method consists of different algorithms, including detection, filtering, validation and fine-tuning. The approach detected over 2,800 potentially hospitalized individuals, yielding a 98.6% agreement with released public records of patients admitted to NHS hospitals. Analyzing the mobility pattern of these individuals prior to their potential hospitalization, we present a series of risk maps. Compared with census-based maps, our risk maps indicate that the areas of highest risk are not necessarily the most densely populated ones. We also show that the areas of highest risk may change from day to day. Finally, we observe that hospitalized individuals tended to have a higher average mobility than non-hospitalized ones. Overall, we conclude that the rich spatio-temporal information extracted from mobile network data may benefit both the mobile-based technologies and the policies that are being developed against existing and future epidemics.

CCS CONCEPTS

• **Networks** → **Mobile networks**; • **Computing methodologies** → **Model development and analysis**; • **General and reference** → **Cross-computing tools and techniques**; **Measurement**; **Evaluation**; **Estimation**; **Validation**.

KEYWORDS

Mobile network data, Signalling data, Human mobility, Epidemic risk map, COVID-19

ACM Reference Format:
Elisa Cabana, Andra Lutu, Enrique Frias-Martinez, and Nikolaos Laoutaris. 2022. Using mobile network data to color epidemic risk maps. In *The 3rd ACM SIGSPATIAL International Workshop on Spatial Computing for Epidemiology (SpatialEpi '22) (SpatialEpi '22), November 1, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3557995.3566120

1 INTRODUCTION

Digital footprints are the records and traces we leave behind us as we use mobiles and the Internet. They often empower tools of significant societal importance, e.g., behavioral models in mental health, civil engineering planning, predicting crime rates, tracking contamination, diseases or viruses [1, 18, 28, 40]. Amidst the COVID-19 epidemic, the necessity to unleash the full potential of digital tools and data-enabled research in the field of health becomes more urgent than ever. From the very first stages of the pandemic many tools have emerged to help with understanding and combating it. One of the most important tools are the *risk maps*, since they visualise the disease distribution and intensity. Coloring areas according to a risk measures is useful for first responders, decision makers, for evaluating the stress on the healthcare system [5], and for decisions making at the level of the individual [49].

1.1 Related work

There are multiple ways to create a risk map, e.g., using census data [15, 47], or cases reported by public health institutions. Another approach based on online surveys has recently been used in several countries [2, 20, 37, 38]. The survey-based methods have the main objective of avoiding the problems that different federal governments or countries had, at least on the initial phases, when evaluating the reach of the epidemic because of limited resources or tests available. Hence, it provides an alternative way of evaluating the number of infections while preserving the privacy of the responses, but it requires a reasonable number of responses which is not always easy to get. The maps based on these high-level statistics may also have accuracy problems because the geographic scope is large, usually country, region or sovereignty [27]. Furthermore, regarding the time dimension, with static maps we could not distinguish whether a certain location is more dangerous during day versus night, or weekday versus weekend.

Static maps are usually based on traditional epidemic models that assume mobility and contact events occur at random. However, reality is more complex as stated by the new type of epidemiology: *the digital epidemiology era* [44], which recognizes the importance of population structure and patterns of interactions and mobility, as elements that can substantially alter the likelihood of disease propagation. In fact, thanks to the ubiquitous cellular devices, the

SpatialEpi '22, November 1, 2022, Seattle, WA, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9543-4/22/11…$15.00
https://doi.org/10.1145/3557995.3566120

study of human mobility has gained significant attention because of the high penetration of cellphones and the low collection cost. To capture this, mobile network data can be effectively exploited to improve our understanding of human mobility dynamics and all social activities and phenomena driven by it [4, 7] including urban planning [30], emergency response [22], and epidemics control [29, 50]. In the latter case, mobile network data has helped to study the risk of infectious diseases contagion [8, 43], the spatial spread of cholera [3], to predict the risk of viruses such as the Zika, malaria or the dengue fever [32, 41, 48, 51], and COVID-19 [17, 19, 26].

An alternative tool is called contact tracing [9, 33], a technique that uses near-field communications and/or GPS at a micro-scale, to study not only the location of a possible interaction that could cause an individual infection or an outbreak, but also the moment in which it occurred. Recent applications based on contact tracing have been developed for the study of the COVID-19 spread, such as the one from Google and Apple [13]. Contact tracing has the advantage of having an increased accuracy since it can identify the interaction at the level of a few meters of distance. However, it suffers from some drawbacks as well, e.g. it requires a large percentage of adoption by the population and it involves serious privacy concerns [6].

1.2 Our Contributions

In this paper, we propose to combine the advantages of the mentioned approaches without suffering from their drawbacks. To achieve this, we developed a method for computing risk maps based on mobile network data that provides detailed spatio-temporal information about millions of cellphones at various scales.

Fig. 1 depicts the three phases of our methodology. In the first phase, we use mobile network data for detecting potential COVID-19 hospitalizations with the detection algorithm that looks for individuals whose phone started appearing during the night in the same area or very close to a hospital that received COVID-19 patients. The result is a set of "pre-detected as hospitalized" individuals. Since false positives and false negatives can occur, we also propose a filtering algorithm in the first phase of the research approach.

The second phase consists of the validation and fine-tuning study based on data released by the National Health Service (NHS), which contains reported COVID-19 hospitalization counts at the level of NHS Trusts. The parameters of the detection and filtering algorithms are fine-tuned towards the combination that gives the best performance with the validation data across different settings.

The third phase starts when the final set of potentially COVID-19 hospitalized individuals is obtained. Then, for each person detected as hospitalized, we study their mobility pattern during the two weeks prior to their day of hospitalization. Based on this, we obtain detailed dynamic risk maps that change through time and thus capture more accurately the distribution, evolution and intensity of the disease.

1.3 Our results

Applying our methodology on a large dataset of more than 2 million cellphones in London, UK, we have identified 2,866 potentially hospitalized COVID-19 cases that compare favorably with released public records that report 11,177 patients admitted to hospitals in London, in the same time frame, which scaled by the 26% mobile networks population coverage is 2,906; thereby yielding a 98.6% agreement with our result. In the validation and fine-tuning study, the parameter configuration that consistently matches the validation data across different settings, while maintaining a high value of correlation between the estimations and the validation data is the cell-tower granularity, with a surrounding area of 750m around the hospitals to choose the cell-towers. The resulting minimum number of consecutive nights that an individual must have spent at the hospital in order to be considered as hospitalized is 4. Analyzing the mobility patterns of the final set of potentially hospitalized individuals, during the two weeks prior to their day of hospitalization, we present a series of risk maps, which show that the multidimensional characteristic of the risk of an area is better reflected when considering spatio-temporal information, as opposed to static data.

1.4 Advantages

The proposed technique has several advantages, for example, (i) it does not require user involvement to collect the data. In fact, the data we use are readily available at various telecommunication companies that use them for operational and other purposes. (ii) Moreover, because this data has a granularity of a base-station antenna or a postcode, they are less privacy-invasive than reading GPS coordinates or near-field contacts of a few meters. Although it is true that mobile network data includes personal information, the fact is that such data are anonymized when used in analytic studies as has been done by many previous works [31], and is also our case. (iii) Furthermore, mobile network data usually contains information about millions of customers, which is an advantage compared to other alternatives such as surveys. (iv) Finally, this proposal allows to have a much better spatial granularity than high-level statistics techniques. We can model mobility using real data in the range of a few hundred meters versus several kilometer granularity extracted from static data, such as census data or infections reported by public health. Our few hundred meters granularity is not enough for extracting contagion probabilities as the contact tracing apps can do, but it is precise enough to detect areas where a high number of positive patients move during different times of the day. Such detailed spatio-temporal characterization of risk cannot be extracted without mobility data.

The paper is organized as follows. Section 2 describes the data used in this research study. Section 3 describes the algorithm for detecting COVID-19 potential hospitalizations from mobile network data, and the parameters involved. Section 4 describes the validation study to test the performance of the proposed approach under different settings and the fine-tuning of the parameters. In Section 5, the algorithm for obtaining risk maps based on the mobility data of the potentially hospitalized individuals, is presented, together with the analysis of the results. Section 6 provides a discussion about the ethical implications, the limitations, and the relevance of the proposed approach for innovative uses of network data beyond communication and real-world impact. Our conclusions are presented in Section 7 together with a description of our ongoing and future work for extending our method.

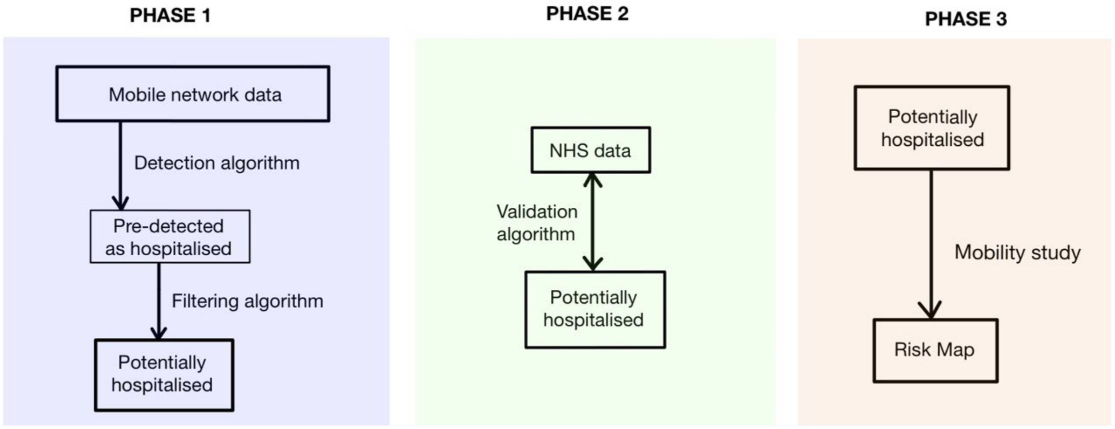

Figure 1: Overview of our research approach.

2 DATA

The data available for this study covers the region of Greater London, United Kingdom. This area has 8.9 millions (2019) inhabitants and spans 1579 square km [16]. In 2019, 95% of the population had a mobile phone of which the vast majority are smartphones [46]. The latter is relevant as the more people connect to the network, the more events are generated and thus the more data points we have per inhabitant. This leads to better mobile network data.

The purpose of a cellular network is to offer mobile communication to subscribers of the system. In order to do this, their location in the network is used to assign them to the most appropriate radio base station. This is done using *signalling data*. With knowledge of the network structure, this data can be transformed to a position of the subscriber. In this paper we used this data to infer the mobility of a set of subscribers through multiple points in time.

The data analyzed in this paper contains the GDPR-compliant and anonymized information of more than 2 million users collected over March and April in 2020, by a British mobile network provider, whose population coverage is approximately of 26%. In our particular case, the records consist of the aggregated user's activity for each night from 00:00h to 08:00h. For each individual we have a date stamp, representing each night. The dataset also contains the top cell-tower from the network, which corresponds to the cell-tower with higher usage time, and the top postcode in which the top cell-tower is included. Finally, it also contains the aggregated duration of the connection events, their home postcode and a one-way hashed id created by the networks provider. Geographic information about the top cell-tower (i.e., their latitude and longitude) was also provided, which in combination with the other features, enable us to study the mobility of individuals at various time frames. The longitudinal aspect of the records is one of the main advantages of mobile network data, since it allows to filter data, and resolve false positives, as we will show.

For the validation and fine-tuning study, we also used a dataset that contains information about the number of hospitalizations reported daily from the start of the epidemic (see [11]). This data is collected on a daily basis and at the level of NHS Trust. More granular information at the level of individual hospitals was not available. In the third phase of the research study we will show how to obtain risk maps. To analyze its results and compare with a baseline, census data from the Office for National Statistics (ONS) in UK, was also used [12]. The UK Census is undertaken every 10 years, with the most recent being on March 27th, 2011. This data provides us with crucial information about the population.

3 DETECTING HOSPITALIZATIONS FROM MOBILE NETWORK DATA

In this section we describe how to exploit mobile network data for detecting potential COVID-19 hospitalizations.

3.1 Formulation

Consider the population of cellphone owners Ω, and a sample $S = \{s_1, s_2, ..., s_n\} \in \Omega$. Let us denote as D_S the mobile network signalling dataset described in Section 2 containing information for each individual $s_i \in S$, $i = 1, ..., n$. Recall that D_S contains the individual's home location h_i, which we have approximated following the methodology in [25]. The dataset D_S also contains daily spatio-temporal information for each individual based on their cellphone activity, such as their top location at night (from 00:00h to 08:00h), i.e., the location in which the person spent most of the time in that period: l_i, and the amount of time spent in that location: t_i.

3.2 Detection algorithm

The objective of the detection algorithm is to perform a binary classification to decide whether each individual in S, can be considered as hospitalized because of COVID-19 or not, by inspecting if their mobile phone appears at night at the same location or near to a hospital admitting COVID-19 patients.

Table 1 shows the parameters involved in the algorithms. One of these parameters is the granularity of the locations, i.e. the categorical parameter φ which takes values $\varphi = \{P, T\}$ depending if the postcode or the cell-tower granularity is chosen, respectively. Therefore, for a fixed individual s_i, both the home location $h_i^{(\varphi)}$ and the top location $l_i^{(\varphi)}$ at night, depend on the two possible levels of granularity φ. In the case of postcodes ($\varphi = P$) the smallest granularity is known as “postcode unit". In the case of cell-towers ($\varphi = T$), it refers to the antenna towers that a user's mobile phone connects to, which, in general, are more granular than the postcodes.

Table 1: Parameters involved in the algorithms.

Concept	Notation	Values	Meaning
Location granularity	φ	P T	Postcode Cell-tower
Cell-tower ratio	r	0.5 0.75 1.0	500 m 750 m 1000 m
Temporal granularity	η	≥ 1	Number of consecutive nights

A set of 74 hospitals in London, admitting COVID-19 patients, was constructed using the information provided in [35]. For each hospital, their location depends on the granularity φ. In the case of postcode level, the location of a hospital is a single value, i.e., the postcode in which that hospital is located. In the case of cell-towers, it is a list of surrounding cell-towers instead of just one, because hospitals are large buildings and individuals inside may connect and disconnect to the several antennas around them, during a certain amount of time. For this purpose, different ratios r around the center of the building are considered, namely 500 meters, 750 meters and 1 kilometer. Now, let us define as $L^{(\varphi,r)}$ the complete list of postcodes or cell-towers, depending on the value of φ, associated to any hospital.

The idea is to cross the information about the individuals' location and activity at night, with hospital locations, and select those individuals that appear at night at a postcode in which a hospital is located, or those connecting to a cell-tower that belongs to the list of towers of some hospital. This could then be used as an indication that an individual may have been hospitalized. Note that this method is generic and configurable, since other criteria could be used, such as home confinement detection instead of hospitalizations. Our approach would probably be more relevant for the initial stages of pandemics, but we are currently studying alternative configurations as described in Section 7.

3.3 Filtering algorithm

The proposed approach can lead to false positives (people detected as hospitalized incorrectly) as well as false negatives (hospitalized people not detected by the approach). Therefore, we formally defined a set of filters to reduce these rates in Table 2.

These filters are shown as conditions in the formulation of the detection and filtering algorithm 1, but let us describe them in detail first and explain how they relate to the false positive or false negative rates. An example of possible false positives could be those individuals that appear at night in the marked locations because they just live in that area, i.e. the s_i for which their home location is in the list of hospital locations: $h_i^{(\varphi)} \in L^{(\varphi,r)}$. Therefore, we set a condition in the algorithm not to include them. We refer to this condition as the “*Home Filter*” (see Table 2). Of course, the opposite can also happen (false negatives), because there might be individuals that live near a hospital and also got sick and hospitalized, but since there is no possible way to detect them, we decided to reduce the false positive rate instead.

Table 2: Formal notation of the filters applied in the algorithm.

Name	Filter condition	Reduction benefit
Home Filter	$h_i^{(\varphi)} \notin L_{\tau_j}^{(\varphi,r)}$	False Positives
Work Filter	$l_i^{(\varphi)} \in L_{\tau_j}^{(\varphi,r)}$ for η nights s.t. $\eta_0 + \eta = \eta_{last}$	False Positives

Another possibility is those that appear at night at the hospital, i.e. their top location at night is in the list of hospital locations: $l_i^{(\varphi)} \in L^{(\varphi,r)}$, but only because they are working there, like nurses, doctors, security guards, etc. It is very common for most workers to show up at the hospital during the night, and even to appear for several consecutive nights (due to night shifts). An individual that appears to be switching location from one night to the other, and thereby exhibiting multiple home-hospital-home-hospital transitions, very rarely will be hospitalized because COVID-19 hospitalizations tend to have a large recovery time period, i.e., a person that has the disease and is hospitalized will typically spend multiple consecutive nights at the hospital [42]. Therefore, individuals located near a hospital but having this alternation behavior will be considered workers from the hospital. We set a condition to filter them out, referred as the “*Work Filter*”. To properly define the latter filter, let us consider a fixed individual s_i that appears at the hospital $l_i^{(\varphi)} \in L^{(\varphi,r)}$ for the total number of η nights. Denote the first night as η_0 and the last night as η_{last}, then this means that this person has stayed at the hospital and did not changed its location in the middle of the hospitalization period. If $\eta_0 + \eta \neq \eta_{last}$, this means that in the middle of the whole period, the individual spent at least one night somewhere else. Although, we could have false negatives with this decision, e.g., workers that get sick and stay hospitalized in that exact same hospital. But since the latter situation would be less frequent, we decided to finally filter out all people potentially working at hospitals. In other words, in the detection step we considered individuals as possibly hospitalized only if they appear at night at the hospital for η consecutive nights, but they neither appear again nor exhibit other transitions.

Moreover, some false positives could also appear for example when people are located in the same area as a hospital at night but not because they are hospitalized, but because they are at a restaurant, hotel, or any other night place that coincidentally is very close to a hospital. In this case, we cannot filter them, but because of the quarantine and the mobility restrictions at the initial stage of the pandemic we believe the false negative rate is low. In fact, two weeks before the start of the government-mandated lockdown on March 23rd, the mobility began to reduce significantly [14, 21, 36]. Furthermore, additional false positives could appear if we decide to consider every person that meets the conditions as COVID-19 hospitalized, while in reality it could be for another cause or another disease. This rate is expected to be low because the time in which the detection is made is the start of the epidemic, and most of the hospital availability was dedicated to COVID-19 patients [23].

In summary, the idea of the algorithm is first filtering out those individuals living in the same location as a hospital. Then, look at their mobile phone activity during the night time and select those users located at night (00:00h-08:00h) at the same area as a hospital for a certain number of consecutive nights. The period selected to perform the detection was April 4-30, 2020, although the epidemic started in London in middle of March. There are two reasons why we choose April, both related to the validation step that we will describe shortly. The first reason is that at the beginning of the pandemic, the reporting of COVID-19 sick and hospitalized people was unreliable, which can ultimately affect our validation process. The second reason is because in London there was a temporary hospital called Nightingale that was created only for COVID-19 patients and it was located at the ExCeL Centre, in an isolated area outside the city. This hospital turned out to be very useful for the validation step. Since it opened in April 4th, we had to perform the detection starting from that date.

Since the COVID-19 virus is very dangerous and hospitalized people are probably under intensive care needs, to consider only one night seems to be very scarce. However, we wanted to study all possibilities, i.e., we explored the values $\eta = 1, 2, 3, ..., T$, where T is the total number of nights in the period of study. In the literature, the average number of consecutive nights that people spend at the hospital, also known as *length of stay*, for COVID-19 disease is set to be at least 4 (see [42]). In the validation study we describe if this makes any sense with respect to our fine-tuning results.

Algorithm 1 depicts the detection and filtering approach for obtaining the final set Φ of detected COVID-19 hospitalized individuals. The approach takes as input the three parameters φ, r, η and dataset D_S containing the aggregated mobile network data of the individuals in S, and geographical information.

```
Input: $D_S, \varphi, r, \eta$
foreach $s_i \in S$ do
    if $h_i^{(\varphi)} \notin L^{(\varphi,r)}$ & $l_i^{(\varphi)} \in L^{(\varphi,r)}$, for $\eta$ nights s.t.
      $\eta_0 + \eta = \eta_{last}$ then
        $s_i \in \Phi$
    else
        $s_i \notin \Phi$
Output: $\Phi$
```

Algorithm 1: Detection and filtering algorithm.

4 VALIDATION AND FINE-TUNING

In the above description of the detection and filtering algorithm, there are some design choices affecting the ratio between false positives and false negatives that cannot be fine-tuned, such as deciding to drop people living in the same area of a hospital in order to benefit the false positive rate. However, as explained above we believe our choices are the ones that best reduce the overall rates. There is a set of three parameters that can be fine-tuned in order to improve the performance of the algorithm, which are φ: the level of granularity, r: the different ratios around the hospitals, and η: the number of consecutive nights at the hospital. The overall objective of the validation and fine-tuning algorithms is to identify the parameter configuration for detecting hospitalizations that matches favorably with the validation data across different settings.

In this study we used the NHS dataset (see Section 2) that contains the number of hospitalizations reported from the start of the epidemic, collected on a daily basis and at the level of NHS Trust. We validate across three different settings: (A) with the total number of patients admitted in hospitals considering all NHS Trust together, (B) with one of the groups called Barts NHS Trust, and (C) with the Nightingale temporary hospital. The reason for the selection of the Barts NHS Trust is because it contains the Nightingale temporary hospital and 6 other hospitals from our list of 74 hospitals in London, admitting COVID-19 patients. We describe now the numerical results, and a summary table is provided at the end of the section.

4.1 Setting A

The reported number of patients admitted to all hospitals is an aggregate of all NHS Trust's number of patients reported in the NHS validation dataset. In the period of April 4-30, 2020, the reported number of patients admitted to hospitals in London was 11,177, which scaled by 26% to capture the mobile network provider's population coverage, is 2,906. In this setting, we apply the algorithm for different values of the parameters, starting with the level of granularity. In the case of tower granularity ($\varphi = T$) we explore three different ratios around the hospitals to obtain the set of cell-towers associated with each one: $r = 500m, 750m$ and $1km$. The last parameter we take into account is η, the minimum number of consecutive nights an individual must have spent at the hospital to be considered as hospitalized. Considering the postcode granularity, the most accurate minimum number of nights was 2. Considering the tower granularity, with the ratio of $1km$ it was at least 4 consecutive nights, with $750m$: 4, and $500m$: 3.

4.2 Setting B

In this setting, only the number of patients admitted to Barts NHS Trust is considered. During the complete period of time, the reported number of hospitalized patients of Barts NHS Trust is 1,098, of which 26% is 285. For the postcode granularity, the most accurate minimum number of nights was 2. Considering the tower granularity, with the ratio of $1km$ it was at least 5 consecutive nights, with $750m$: 4, and $500m$: 3. Note that the parameter sets ($\{\varphi = T, r = 0.75, \eta = 4\}$ and $\{\varphi = T, r = 0.5, \eta = 3\}$) that provide the best match between predicted and actual hospitalizations is the same with the corresponding parameter set of setting A.

4.3 Setting C

In this last setting we consider the Nightingale temporary hospital. Since the validation data does not contain any individual hospital information, the real number of patients admitted in Nightingale at its individual level was not provided. This information was found in other sources from the press such as [34]. The total number of reported hospitalizations in this hospital was 51, therefore scaling by 26% gives 13. The results indicate that the most accurate minimum number of nights, with postcode granularity was 11 nights, and with tower granularity, it was: 10 ($1km$), 4 ($750m$), and 4 ($500m$).

The parameter set ($\{\varphi = T, r = 0.75, \eta = 4\}$) providing the best match between predicted and actual hospitalizations was the same as the corresponding parameter set of setting A and B.

4.4 Correlation study

For each setting, the daily robust Spearman correlation coefficient ρ between the algorithm results and the reported daily cases in the validation data, is obtained. In the first setting A, where all hospitals are considered, depending on the ratios and the minimum number of nights, the correlation results were: 1km ($\geq$ 4 nights): 0.885, 750m ($\geq$ 4 nights): 0.896, 500m ($\geq$ 3 nights): 0.826. All the correlations are high, the maximum in the case when the $750m$ ratio is chosen. In setting B, where the Bart NHS Trust is studied, the resulting daily robust correlation coefficient is, for 1km ($\geq$ 5 nights): 0.801, 750m ($\geq$ 4 nights): 0.862, 500m ($\geq$ 3 nights): 0.638. Again, they all are high values, but the maximum number is when the $750m$ ratio is chosen. Daily correlation in setting C, for Nightingale temporary hospital cannot be computed because we only have total cases information from this hospital in the overall period, not daily reported cases.

4.5 Summary of results

Table 3 summarizes the results in the validation and fine-tuning study that we described above. Note that the parameter configuration that consistently matches the validation data across different validation settings, while maintaining a high value of correlation ρ between the estimations and the validation data (marked in bold), is $\{\varphi = T, r = 0.75, \eta = 4\}$, which means that the minimum number of consecutive nights is selected to be at least 4, the granularity in the detection is set to be at the cell-tower level, and the area around the hospitals to choose the towers is set to be of 750 meters.

Table 3: Results in the validation study.

Setting	Description	φ, r	Min. η	ρ
A	All Hospitals	$\varphi = P$	2	0.623
		$\varphi = T, r = 0.5$	4	0.885
		$\boldsymbol{\varphi = T, r = 0.75}$	**4**	**0.896**
		$\varphi = T, r = 1$	3	0.826
B	Barts NHS Trust	$\varphi = P$	2	0.637
		$\varphi = T, r = 0.5$	5	0.801
		$\boldsymbol{\varphi = T, r = 0.75}$	**4**	**0.862**
		$\varphi = T, r = 1$	3	0.638
C	Nightingale Hospital	$\varphi = P$	11	X
		$\varphi = T, r = 0.5$	10	X
		$\boldsymbol{\varphi = T, r = 0.75}$	**4**	X
		$\varphi = T, r = 1$	4	X

5 RISK MAPS

In the previous sections we described our approach for detecting potential hospitalizations using mobile network data. In this section we describe how diverse risk maps based on this data can be obtained. The main idea is to explore the mobility of the individuals detected as hospitalized, during the period of two weeks prior to their first day of hospitalization. The two weeks duration was selected because this was the maximum estimated incubation period of COVID-19, at least at the initial stage of the epidemic [39], in which there were no variants of the virus and people were not yet vaccinated. The approximate spatio-temporal trajectory of a mobile phone and its user can be reconstructed by linking the mobile network data associated with that phone with the geographic location of the cellular tower, or the postcode containing that tower.

5.1 Algorithm for obtaining the risk maps

To create a risk map, two things must be decided. First, the granularity, and second, the measure of risk. The granularity, that we define as the parameter λ, can be postcode levels (units, sectors, or districts), or census levels (OA, LSOA, MSOA, or Boroughs). Once a granularity is decided, a time-lapse choropleth map can be obtained, which can be turned into a movie depicting the evolution of the map through time. Unfortunately, we were not be able to obtain the mobility data for the whole day, but only for the aggregated eight hours at night from 00:00-08:00. In the future, when this data is available we can apply the same method to obtain a complete daily mobility based risk map, but for now we believe it is important to show the potential advantages of this approach, even with reduced data, since our work can motivate further research in the area.

The measure of risk, in our case, depends on the number of people detected as hospitalized that are at the same time in the same location, two weeks prior to the moment of their hospitalization, i.e., it can be seen as an 'a-priori' measure of risk. A higher number of people located at the same time in a fixed area increases the risk of infection for everyone else, due to exposure [10, 24]. This measure of risk could also depend on the time spend by each person in that location, i.e., the longer an infected person is in a fixed location, the more risk there is. In fact, from epidemiology models, several different measures of risk can be computed, but supporting which is the best epidemiology model is beyond the scope of our work. When a particular measure of risk is computed, it can be considered as a weight for each specific area and based on that, a color can be assigned to each area and a risk map can be obtained.

As depicted in the Algorithm 2 for obtaining the risk maps, for each granularity and each day, we can compute different individual risk maps, since the risk measures vary through time and space. A *risk map movie*[1] can be obtained if we consider the time-lapse set of daily risk maps for a fixed granularity. This result can help capturing the evolution of the epidemic spread pattern and the spatial risk of contagion. The parameters involved in Algorithm 2 are the following:

- T is the total number of days in the complete period of time.
- $t = 1, ..., T$ is the daily time steps.
- λ is the parameter that selects the geography to plot the risk map (e.g., Borough, Postcode District, etc.).
- D_λ is the dataset containing the polygon geometries for plotting the map depending on the selected λ.
- A is the total number of areas in the map.
- $l_k \in D_\lambda$, for $k = 1, ..., A$, are the location areas in the map.
- $p^{(l_k,t)}$ is the measure of risk of area l_k at time t, that depends on the number of persons detected as hospitalized ($s_i \in \Phi$)

[1] An example of risk map movie can be found here.

that are at the same time in the same location, in the period of two weeks prior to their hospitalization.

- $a_i^{(l_k,t)}$, for $i = 1, \ldots, |\Phi|$, counts the number of users (detected as hospitalized) located in area l_k at time t.

```
Input: $\Phi$, $\lambda$, $D_\lambda$
foreach $t = 1, \ldots, T$ do
    foreach $k = 1, \ldots, A$ do
        foreach $i = 1, \ldots, |\Phi|$ do
            if $s_i$ is in area $l_k$ at time $t$ then
                $a_i^{(l_k,t)} = 1$
            else
                $a_i^{(l_k,t)} = 0$
        $p^{(l_k,t)} = \sum_{s_i \in \Phi} a_i^{(l_k,t)}$
    Risk Map $(\lambda, t)$
Output: Risk Map Movie $(\lambda)$
```

Algorithm 2: Algorithm for obtaining the risk maps.

5.2 Comparison with static maps

We obtained various risk map movies at different granularities. Figures 2 and 3 show, as examples, the average risk map versus the static census map, at Borough and District levels, respectively. The color range in the risk map, from white to purple (less to higher), means that the more purple an area, the more dangerous (in case of a risk map) it is. Areas with higher weight, are those where more sick people appeared there or spent time. From these risk maps we can obtain statistical information to gain further insight about the epidemic evolution. In average, the top 10 most dangerous areas at the Borough level were: Westminster, Newham, Lambeth, Tower Hamlets, Camden, Kensington and Chelsea, Southwark, Hackney, Islington and Brent. At the District level they were: E1, NW8, SE1, E6, W10, SW9, NW1, W2, N19 and SW17. Although interpreting one by one the demographics and the reasons that the top 10 areas appear to be the riskiest ones is beyond the scope of this paper, we note that most of them are well known areas with intense night life due to the presence of restaurants, bars, etc. Let us now compare our risk maps with the static population map resulting from the census data [12], and let us compute the correlations between the two types of maps.

Variability in the spatial dimension: Figures 2 and 3 show the static census-based population maps (down), for Boroughs and Districts, respectively. As it can be seen at first sight, the average risk maps (up) look very different in comparison with their respective analogous population map (down). In fact, if we look at the risk map in Fig. 2 we see that the most dangerous borough is Westminster, which has the highest value on April 24th (Friday). Meanwhile, if we look at the census map in Fig. 2, we see that Croydon is actually the borough with the highest population and it is not in the top 10 most dangerous boroughs. Moreover, the most dangerous postcode district is E1, in average and most of the times. The highest value in E1 is also on April 24th, 2020. Although, according to the census data, CR0 is the district with the highest population density, and it was not found within the top 10 most dangerous districts.

Variability in the time dimension: Not only do the risk maps differ, in average, from the corresponding population map, but if we add the time dimension and consider the different dynamic risk maps for each night, this disagreement increases. To measure

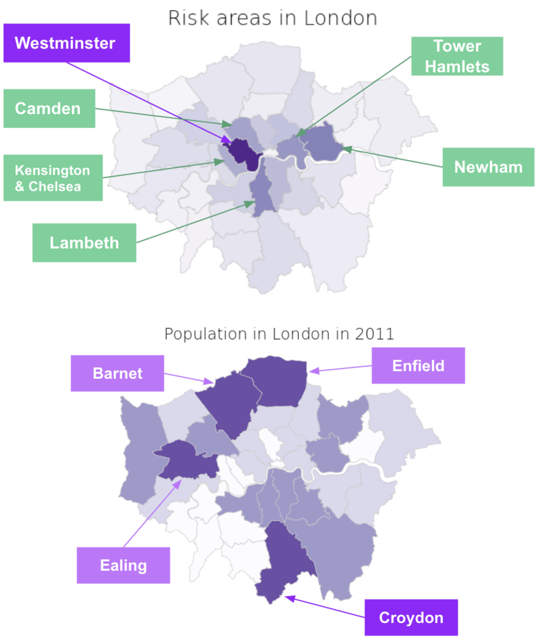

Figure 2: Risk map (up) versus Census map (down), at borough level.

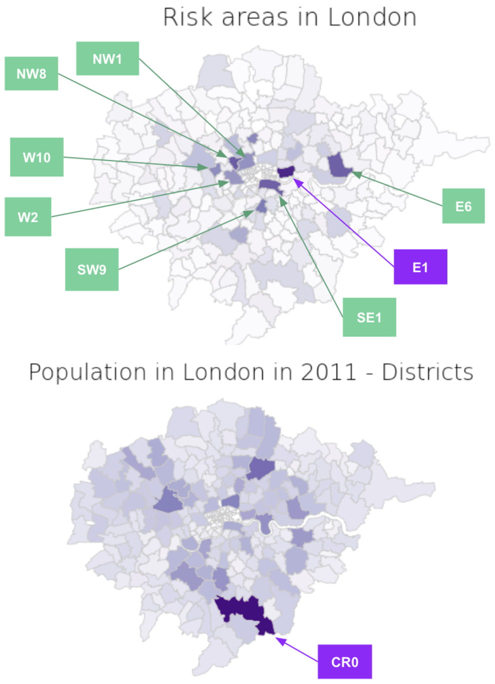

Figure 3: Risk map (up) versus Census map (down), at district level.

these differences numerically, for each granularity level, we also computed the correlations between the daily risk maps and the population map. The average correlation and standard deviation are shown in Table 4, for each granularity level.

Table 4: Results of the correlation study between the risk map and the population map.

Level	Average correlation	Std. Dev.
Borough	0.1462	0.0242
District	0.0959	0.0275

The resulting low correlations indicate that the risk maps could not be inferred by simply taking into account the population density, i.e., the areas of highest risk are not necessarily the most densely populated ones, but a result of multiple factors which may be better reflected when taking the spatio-temporal information of higher granularity like the one proposed in this paper, as opposed to static data like census.

Moreover, there is also some variability within the group of individual daily risk maps and, more concretely, the areas of highest risk vary from day to day. Additionally, if we use the time dimension to compute what are the most dangerous days for the top 10 (in average) most dangerous areas, we discover that most times, these are weekends. In fact, if we consider the average between all the top 10 most dangerous areas, the top 1 most dangerous day is April 3rd (Friday). And the top 2 and 3 most dangerous days are also from early April, i.e., the 4th (Saturday) and the 5th (Sunday).

5.3 Bounding box analysis

The risk maps can be computed aggregating data to the desired level of granularity, e.g., postcode or census levels. However, note that the available data is highly granular, and we have information about the location of the individuals at the level of cell-towers. That means that additional mobility studies can be performed, for example, computing for each user the bounding box that contains the set of locations on which the user moves. Then, for each individual, we can compute the area or the diagonal of their bounding box, or any other measure that aims to represent the extent of their mobility. In our case, for the (detected as) hospitalized individuals, the average diagonal of their bounding box was 8.5km, with a standard deviation of 10.27km. Meanwhile, for the rest of individuals, the average diagonal of their bounding box was 5.29km, with a standard deviation of 5.82km. We performed a hypothesis test comparing the average diagonals of the bounding boxes for each type of individual, and with significance level of $\alpha = 0.05$ the results were statistically significant. The test statistic is 16.61, much higher than the critical point $z_{\alpha=0.05} = 1.645$, which results in rejecting the null hypothesis, i.e., with 95% of confidence, the data favors the alternative hypothesis that the hospitalized users have an average mobility higher than the ones not detected as hospitalized.

As stated in the introduction, several approaches can be used to obtain a risk map. However, most of them do not take into account the spatio-temporal dynamics of the individual's behaviour. As we could see in the comparison against the census maps, these are static outputs resulting from long-time studies (usually static for years), describing a single characteristic per map, e.g., population, housing, income, retail, etc. Meanwhile, an epidemic spread is of multidimensional nature, and mobility of individuals could explain this multiple factor dependant event, at once. Other maps for epidemics could be obtained based on the number of cases or deaths in certain areas (either reported by health or governmental institutions, or obtained through surveys), although, they cannot depict risk based on mobility and they are usually obtained for complete days, i.e., they cannot be plotted hourly nor in real time. Epidemic risk maps based on mobile network data have the potential to describe the epidemic spread and evolution in different and more granular spatio-temporal settings.

6 DISCUSSION

In this section we discuss the ethical implications, the limitations, and the relevance of the proposed approach for innovative uses of network data beyond communication and for impact in the real world.

6.1 Ethical implications

As stated in Section 2, the mobile network data used in this paper has been reviewed and validated by the operator with respect to GDPR compliance (e.g., no identifier can be associated to person). A limitation is that the data do not have explicit user consent for these studies, although it has been used in previous analytic studies, such as [31]. Data collection and retention at the mobile network operator are in accordance with their terms and conditions and the local regulations. Any re-sharing of the data with 3rd parties even for research purposes is prohibited. No personal and/or contract information was available for this study and none of the authors of this paper participated in the extraction and/or encryption of the raw data. The temporal resolution of mobile network data, reflected in the timestamps of the user's activity in this data, is at the aggregated eight hours level. With respect to the algorithm, it is applied directly to the anonymized data, and the results cannot leak any private information. Furthermore, we have successfully obtained ethical approval for this research from our institutions.

6.2 Limitations

The proposed approach aims to detect potential hospitalizations and risk maps based on past mobility information. A limitation is that epidemic risk is not only based on severe cases, but also depends on the asymptomatic cases spreading the disease. It is in fact a complex matter and an interplay between many factors. However, the proposed approach could be relevant in future epidemic outbreaks, since in the beginning very little is usually known about new epidemic viruses. Furthermore, mobile network data is always there at disposal, ready to be studied and transformed into models and tools that can give us valuable insights on the extent, spread and evolution of the outbreak. We are currently studying how to expand our approach, to detect potential COVID-19 positives that could be more relevant now, e.g., people that have mild symptoms and stay at home doing quarantine.

6.3 Innovation and real-world impact

The ubiquity of mobile devices and the unique advantages of cellular networks (e.g., availability and high penetration rates) have facilitated an increase in the innovative uses of network data beyond communication, such as the several large-scale human mobility research studies based on mobile network data. One of the main barriers for organizations to adopt existing technologies, such as contact tracing apps, is the complexity of usage along with their requirements. For example, the exposure notification apps developed by Apple and Google with Bluetooth data were not easy to be adopted because they required digital certificates and having users to install an application was an obstacle. Meanwhile, mobile network data, in general, has two major strengths: (i) high penetration rates, i.e., most people take cellphones with them every day; (ii) no need for incentives and low collection cost, i.e., the network providers already have this data ready for further research and no extra cost or additional incentives need to be introduced for data collection. In fact, the *mobile network signalling data* that is used in this paper has the additional advantage of allowing to capture the mobility of the users all the time, not just when they generate active data, e.g., voice or SMS communications.

Our approach is the first step in the direction of using mobile network data to derive equivalent estimations without the need of healthcare providers' interventions, digital certificates or user adoption. The proposed approach is generic and configurable, and it can be applied to different cities or countries, depending on the data available. Future methods can be developed based on this idea and telecommunication companies could implement very easily fully automated services to, e.g., estimate the dangerous areas or risky contacts and send an SMS to warn individuals when they enter a high risk area as a reminder to take extra protective measures.

Furthermore, our algorithm for detecting COVID-19 hospitalizations can be interpreted as a severity detection approach which could also have high real-world impact if used for identifying healthcare system burden due to an increased number of hospitalizations. Moreover, computing a measure of risk of an area experiencing an outbreak, can be very helpful for society because it can help individuals take extra protective measures, as well as first responders and decision makers in evaluating the spread of severity, the healthcare system burden, the 'a-priori' higher risk areas where people potentially got infected, and the advantages and disadvantages of alternative courses of action. Additionally, this is closely related to spatial hotspot mapping, i.e., finding regions with significantly higher rates of generating cases of certain events. In this context, a risk measure can be used to detect future emerging high risk regions using forecast methods with time series data [45, 52].

Overall, we conclude that the rich spatio-temporal information extracted from mobile network data may benefit both the mobile technologies and the policies that are being developed against existing and future epidemics.

7 CONCLUSIONS AND FUTURE WORK

In this paper we propose an approach to detect potential COVID-19 hospitalizations and epidemic risk maps, based on mobile network data containing detailed spatio-temporal information about millions of cellphones at various scales. The methodology consists of three phases: (i) detecting potential COVID-19 hospitalizations, (ii) performing a validation and fine-tuning study, and (iii) analyzing the mobility patterns of the final set of hospitalized individuals, prior to their hospitalization, for obtaining a series of risk maps.

The approach detected 2,866 potentially hospitalized individuals, yielding a 98.6% agreement with released public records of patients admitted to hospitals in London, in the same time frame. We compare our proposed risk maps with static census-based maps and the results show that the areas of highest risk are not necessarily the most densely populated ones. This disagreement increases when we add the time dimension and consider the different dynamic risk maps for each night versus the static map. We computed the top 10 most dangerous areas (in average) at different granularity levels (e.g., postcode districts and boroughs) and we note that most of them are well known areas with intense night life due to the presence of several tourist places, restaurants, bars, etc. We also used the time dimension to compute the most dangerous days for the top 10 most dangerous areas and we discovered that most times these are weekends. We also found that the areas of highest risk vary from day to day. Finally, we found that the individuals detected as hospitalized by our approach, have had a statistically significant higher average mobility than the ones not detected as hospitalized. The conclusion is that the multidimensional characteristic of the risk of an area is better reflected when taking spatio-temporal information of high granularity like the one proposed in this paper.

For future work, with the complete data containing the mobility information during the whole day, we plan to have a more general vision of how the mobility of people influences the risk. We will also be able to study a more granular time dimension (e.g., hourly) and time-dependent similarities or differences (e.g., weekday vs weekend, or between morning, noon, afternoon, evening and night). Furthermore, once we manage to compute a measure of risk for each area and time interval, we could use time series analysis to fit a model that estimates the future risk for each area. An exogenous variable could be the number of other healthy people found in the same area at that exact time, which can influence (by auto-correlations) the number of sick people and can be entered as extra information to the model in the moment of forecasting the next future day (assuming we know how many people are going to be there). Our proposed approach is generic and can be configured using different criteria, such as the one we are currently studying that consists of detecting home confinements and fine-tuning other parameters like the incubation period. Finally, combining our model and a contagion risk alert system like the one proposed for H1N1 virus by [18], we could also provide an alert system based on the identification of risky physical contacts.

ACKNOWLEDGMENTS

This paper has been funded by the Regional Government of Madrid and the European Union through the European Regional Development Fund (ERDF) as part of the response from the European Union to the COVID-19 pandemic in the context of REACT-CONTACT-CM-23479 research project, awarded under the Agreement formalized between the Regional Government of Madrid and IMDEA Networks on April 19th 2022.

REFERENCES

[1] R. Agarwal and A. Banerjee. 2020. Infection Risk Score: Identifying the risk of infection propagation based on human contact. In *ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19*.

[2] J. Álvarez, C. Baquero, E. Cabana, J. P. Champati, A. F. Anta, D. Frey, A. García-Agúndez, C. Georgiou, M. Goessens, H. Hernández, et al. 2021. Estimating Active Cases of COVID-19. *arXiv preprint arXiv:2108.03284*.

[3] L. Bengtsson, J. Gaudart, X. Lu, S. Moore, E. Wetter, K. Sallah, S. Rebaudet, and R. Piarroux. 2015. Using mobile phone data to predict the spatial spread of cholera. *Scientific reports*.

[4] V. D. Blondel, A. Decuyper, and G. Krings. 2015. A survey of results on mobile phone datasets analysis. *EPJ data science*.

[5] G. Bobashev, I. Segovia-Dominguez, Y. R. Gel, J. Rineer, S. Rhea, and H. Sui. 2020. Geospatial forecasting of COVID-19 spread and risk of reaching hospital capacity. *SIGSPATIAL Special*.

[6] L. Bradford, M. Aboy, and K. Liddell. 2020. COVID-19 contact tracing apps: a stress test for privacy, the GDPR, and data protection regimes. *Journal of Law and the Biosciences*.

[7] F. Calabrese, L. Ferrari, and V. D. Blondel. 2014. Urban sensing using mobile phone network data: a survey of research. *ACM CSUR*.

[8] M. A. Carrillo, A. Kroeger, R. C. Sanchez, S. D. Monsalve, and S. Runge-Ranzinger. 2021. The use of mobile phones for the prevention and control of arboviral diseases: a scoping review. *BMC public health*.

[9] M. Cebrian. 2021. The past, present and future of digital contact tracing. *Nature Electronics*.

[10] D. Crichton. 1999. The risk triangle. *Natural disaster management*.

[11] NHS Dataset. 2021. National Health Service (NHS) in England. Retrieved June 20, 2021 from https://www.england.nhs.uk/statistics/statistical-work-areas/covid-19-hospital-activity/

[12] ONS Population Dataset. 2021. Office for National Statistics. Retrieved June 20, 2021 from http://www.ons.gov.uk

[13] S. Davalbhakta, S. Advani, S. Kumar, V. Agarwal, S. Bhoyar, E. Fedirko, D. P. Misra, A. Goel, and L. Gupta. 2020. A systematic review of smartphone applications available for corona virus disease 2019 (COVID19) and the assessment of their quality using the mobile application rating scale (MARS). *Journal of medical systems*.

[14] W. Do Lee, M. Qian, and T. Schwanen. 2021. The association between socioeconomic status and mobility reductions in the early stage of England's COVID-19 epidemic. *Health & Place*.

[15] E. Dong, H. Du, and L. Gardner. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*.

[16] Eurostat. 2021. Eurostat Databases. Retrieved June 20, 2021 from https://ec.europa.eu/eurostat/data/database

[17] Z. Fan, X. Song, Y. Liu, Z. Zhang, C. Yang, Q. Chen, R. Jiang, and R. Shibasaki. 2020. Human mobility based individual-level epidemic simulation platform. *SIGSPATIAL Special*.

[18] E. Frıas-Martınez, G. Williamson, and V. Frıas-Martınez. 2013. Simulation of epidemic spread using cell-phone call data: H1N1 case study. In *Netmob'13*.

[19] S. Gao, J. Rao, Y. Kang, Y. Liang, and J. Kruse. 2020. Mapping county-level mobility pattern changes in the United States in response to COVID-19. *SIGSPATIAL Special*.

[20] A. Garcia-Agundez, O. Ojo, H. Hernandez, C. Baquero, D. Frey, C. Georgiou, M. Goessens, R. E. Lillo, R. Menezes, N. Nicolaou, et al. 2021. Estimating the COVID-19 Prevalence in Spain with Indirect Reporting via Open Surveys. *Frontiers in Public Health*.

[21] Google. 2021. Google COVID-19 community mobility reports. Retrieved June 20, 2021 from https://www.google.com/covid19/mobility/

[22] D. Gundogdu, O. Incel, A. Salah, and B. Lepri. 2016. Countrywide arrhythmia: emergency event detection using mobile phone data. *EPJ Data Science*.

[23] J. A. Hardie and P. A. Brennan. 2020. Are you surgically current? Lessons from aviation for returning to non-urgent surgery following COVID-19. *British Journal of Oral and Maxillofacial Surgery*.

[24] S. Hazarie, D. Soriano-Paños, A. Arenas, J. Gómez-Gardeñes, and G. Ghoshal. 2021. Interplay between population density and mobility in determining the spread of epidemics in cities. *Communications Physics*.

[25] S. Isaacman, V. Frias-Martinez, and E. Frias-Martinez. 2018. Modeling human migration patterns during drought conditions in La Guajira, Colombia. In *ACM SIGCAS COMPASS*.

[26] J. S. Jia, X. Lu, Y. Yuan, G. Xu, J. Jia, and N. A. Christakis. 2020. Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature*.

[27] M. Kiamari, G. Ramachandran, Q. Nguyen, E. Pereira, J. Holm, and B. Krishnamachari. 2020. COVID-19 Risk Estimation using a Time-varying SIR-model. In *ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19*.

[28] N. E. Kogan, L. Clemente, P. Liautaud, J. Kaashoek, N. B. Link, A. T. Nguyen, F. S. Lu, P. Huybers, B. Resch, C. Havas, et al. 2021. An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time. *Science Advances*.

[29] A. Lima, M. De Domenico, V. Pejovic, and M. Musolesi. 2015. Disease containment strategies based on mobility and information dissemination. *Scientific reports*.

[30] T. Louail, M. Lenormand, O. G. C. Ros, M. Picornell, R. Herranz, E. Frias-Martinez, J. J. Ramasco, and M. Barthelemy. 2014. From mobile phone data to the spatial structure of cities. *Scientific reports*.

[31] A. Lutu, D. Perino, M. Bagnulo, E. Frias-Martinez, and J. Khangosstar. 2020. A characterization of the COVID-19 pandemic impact on a mobile network operator traffic. In *Proceedings of the ACM internet measurement conference*.

[32] L. Mao, L. Yin, X. Song, and S. Mei. 2016. Mapping intra-urban transmission risk of dengue fever with big hourly cellphone data. *Acta tropica*.

[33] M. Mokbel, S. Abbar, and R. Stanojevic. 2020. Contact tracing: Beyond the apps. *SIGSPATIAL Special*.

[34] Press News. 2021. Nightingale Temporary Hospital press news. Retrieved April 20, 2021 from https://www.bmj.com/content/369/bmj.m1860

[35] NHS. 2021. NHS Trusts. Retrieved June 20, 2021 from https://www.nhs.uk/

[36] P. Nouvellet, S. Bhatia, A. Cori, K. Ainslie, M. Baguelin, S. Bhatt, A. Boonyasiri, N. F. Brazeau, L. Cattarino, L. V. Cooper, et al. 2021. Reduction in mobility and COVID-19 transmission. *Nature communications*.

[37] O. Ojo, A. García-Agundez, B. Girault, H. Hernández, E. Cabana, A. García-García, P. Arabshahi, C. Baquero, P. Casari, E. J. Ferreira, et al. 2020. CoronaSurveys: Using Surveys with Indirect Reporting to Estimate the Incidence and Evolution of Epidemics. *arXiv preprint arXiv:2005.12783*.

[38] N. Oliver, J. X. Barber, K. Roomp, and K. Roomp. 2020. Assessing the Impact of the COVID-19 Pandemic in Spain: Large-Scale, Online, Self-Reported Population Survey. *Journal of medical Internet research*.

[39] World Health Organization. 2021. Coronavirus disease (COVID-19). Retrieved June 20, 2021 from https://www.who.int/health-topics/coronavirus#tab=tab_3

[40] U. Qazi, M. Imran, and F. Ofli. 2020. GeoCoV19: a dataset of hundreds of millions of multilingual COVID-19 tweets with location information. *SIGSPATIAL Special*.

[41] J. Rajarethinam, J. Ong, S. Lim, Y. Tay, W. Bounliphone, C. Chong, G. Yap, and L. Ng. 2019. Using human movement data to identify potential areas of Zika transmission: case study of the largest Zika cluster in Singapore. *International journal of environmental research and public health*.

[42] E. M. Rees, E. S. Nightingale, Y. Jafari, N. R. Waterlow, S. Clifford, C. A. B. Pearson, T. Jombart, S. R. Procter, G. M. Knight, CMMID Working Group, et al. 2020. COVID-19 length of hospital stay: a systematic review and data synthesis. *BMC medicine*.

[43] S. Rubrichi, Z. Smoreda, and M. Musolesi. 2018. A comparison of spatial-based targeted disease mitigation strategies using mobile phone data. *EPJ Data Science*.

[44] M. Salathe, L. Bengtsson, T. J. Bodnar, D. D. Brewer, J. S. Brownstein, C. Buckee, E. M. Campbell, C. Cattuto, S. Khandelwal, P. L. Mabry, et al. 2012. Digital epidemiology. *PLoS Comput Biol*.

[45] R. Souza, R. Assunção, D. Neill, and W. Meira Jr. 2019. Detecting spatial clusters of disease infection risk using sparsely sampled social media mobility patterns. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*.

[46] Statista. 2021. Statista. Retrieved June 20, 2021 from https://www.statista.com/

[47] Z. Sun, L. Di, W. Sprigg, D. Tong, and M. Casal. 2020. Community venue exposure risk estimator for the COVID-19 pandemic. *Health & Place*.

[48] A. J. Tatem, Z. Huang, C. Narib, U. Kumar, D. Kandula, D. K. Pindolia, D. L. Smith, J. M. Cohen, B. Graupe, P. Uusiku, et al. 2014. Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning. *Malaria journal*.

[49] G. Thakur, K. Sparks, A. Berres, V. Tansakul, S. Chinthavali, M. Whitehead, E. Schmidt, H. Xu, J. Fan, D. Spears, et al. 2020. COVID-19 joint pandemic modeling and analysis platform. In *ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19*.

[50] M. Tizzoni, P. Bajardi, A. Decuyper, G. K. K. King, C. M. Schneider, V. Blondel, Z. Smoreda, M. C. González, and V. Colizza. 2014. On the use of human mobility proxies for modeling epidemics. *PLoS Comput Biol*.

[51] A. Wesolowski, T. Qureshi, M. F. Boni, Pål R. Sundsøy, M. A. Johansson, S. B. Rasheed, K. Engø-Monsen, and C. O. Buckee. 2015. Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proceedings of the National Academy of Sciences*.

[52] Y. Xie, S. Shekhar, and Y. Li. 2022. Statistically-robust clustering techniques for mapping spatial hotspots: A survey. *Comput. Surveys* (2022).