Improving epidemic risk maps using mobility information from mobile network data

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 by a mobile network provider located in London, UK. The approach yields 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 and may change from day to day. Finally, we observe that hospitalized individuals tended to have a higher average mobility than non-hospitalized ones.

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

1 INTRODUCTION

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 [3], and for decisions making at the level of the individual [22].

1.1 Related work

There are multiple ways to create a risk map, e.g., using census data [10, 21], or cases reported by public health institutions. Another approach based on online surveys has recently been used in several countries [1, 18]. The survey-based methods provide an alternative way of evaluating the number of infections while preserving the privacy of the responses but require a reasonable number of responses, which is not always easy to get.

Static maps are usually based on traditional epidemic models that assume that mobility is a random process. However, the population structure, patterns of interactions and mobility, are elements that can substantially alter the likelihood of disease propagation. To capture this, mobile network data can be effectively exploited to improve our understanding of human mobility dynamics and its impact on infectious diseases contagion [5], the spatial spread of cholera [2]. Also to predict the risk of viruses such as the Zika, malaria and the dengue fever [19], and COVID-19 [11, 12, 15].

An alternative tool is Contact Tracing [6, 16], a technique that uses near-field communications and/or GPS at a micro-scale. 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 [9]. 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 since it requires a large percentage of adoption by the population and it involves serious privacy concerns [4].

1.2 Our Contributions

In this paper we propose a method for computing risk maps based on mobile network data that provides detailed spatio-temporal information about millions of cellphones at various scales. The proposed method consists of three phases. In the first phase (Section 2), we describe the algorithm for detecting COVID-19 potential hospitalizations from mobile network data, and the parameters involved. The second phase (Section 3) consists of the validation and fine-tuning study based on data released by the National Health Service (NHS). In the third phase (Section 4), 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. Our conclusions are presented in Section 5.

2 DETECTING HOSPITALIZATIONS FROM MOBILE NETWORK DATA

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

2.1 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* containing information for each individual $s_i \in S$, $i = 1, ..., n$. This data covers the region of Greater London, United Kingdom and 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. D_S contains the individual's home location h_i, which has been approximated following the methodology in [14]. 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 area: t_i.

2.2 Detection

The objective is to obtain the set of potentially hospitalized individuals 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 algorithm.

Table 1: Parameters involved in the algorithm.

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
Time granularity	η	≥ 1	# of consecutive nights

A set of 74 hospitals in London, admitting COVID-19 patients, was constructed using the information provided in [17]. 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.

2.3 Filters

The approach can lead to false positives (incorrectly detected as hospitalized) as well as false negatives (true hospitalized not detected). Therefore, we defined a set of filters to reduce these rates.

Home Filter: Individuals can appear at night in the marked locations because they just live in that area. These are potential false positives. To exclude them we set a condition to consider only the subscribers s_i for which their home location is not in the list of hospital locations: $h_i^{(\varphi)} \notin L^{(\varphi,r)}$.

Work Filter: Individuals can appear at night at the hospital because they are working there, e.g. nurses, doctors, security guards, etc. If they exhibit multiple home-hospital transitions, very rarely they will be hospitalized due to COVID-19, because this disease tends to have a large recovery time period [20]. Therefore, 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}. If $\eta_0 + \eta = \eta_{last}$, this means that this person has stayed at the hospital and did not changed its location in the middle of the hospitalization period. If not, this means that in the middle of the whole period, the individual spent at least one night somewhere else. Then, we set a condition to consider only the individuals that appear at night at the hospital for η consecutive nights, but they neither appear again nor exhibit other transitions.

2.4 Detection and Filtering algorithm

Algorithm 1 depicts the Detection and Filtering approach for obtaining the final set Φ of detected COVID-19 hospitalized individuals.

```
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.

3 VALIDATION AND FINE-TUNING

The parameters that can be fine-tuned in order to improve the algorithm's performance 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 is to identify the parameter configuration that matches favorably with the validation data across different settings.

In this study we used the NHS dataset containing the number of daily hospitalizations reported from the start of the epidemic, 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. For each setting, the daily robust Spearman correlation coefficient ρ between the algorithm results and the reported daily cases in the validation data, is also obtained. Daily correlation in setting C, for Nightingale temporary hospital cannot be computed because daily reported cases for this hospital was not available.

Table 2 summarizes the results. Note that the parameter configuration that consistently matches the validation data across different settings, while maintaining a high value of correlation ρ

between the estimations and the validation data (marked in bold) is $\{\varphi = T, r = 0.75, \eta = 4\}$. This configuration yields a 98.6% agreement with released public records of patients admitted to hospitals in London, within the same time frame.

Table 2: 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

4 RISK MAPS

In this section we describe how diverse risk maps can be obtained to explore the mobility of the individuals detected as hospitalized, during the period of two weeks prior to their first day of hospitalization. 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.

4.1 Algorithm for obtaining the risk maps

We propose to define a measure of risk that 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 [7, 13]. 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.

Algorithm 2 describes our proposed method for obtaining the daily risk maps. 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: the total number of days.
- $t = 1, \ldots, T$: the daily time steps.
- λ: the geography to plot the risk map.

```
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.

- D_λ: the polygon geometries depending on the selected λ.
- A: the total number of areas in the map.
- $l_k \in D_\lambda$ $(k = 1, \ldots, A)$: the location areas in the map.
- $a_i^{(l_k,t)}$, for $i = 1, \ldots, |\Phi|$: counts the number of users (detected as hospitalized) located in area l_k at time t.
- $p^{(l_k,t)}$: the measure of risk of area l_k at time t.

4.2 Comparison with static maps

Various risk map movies at different granularities (OA, LSOA, MSOA, Boroughs, postcode units, sectors or districts) were obtained. Figure 1 shows the average risk map versus the static census map in the case of Borough level granularity, as an example. The color range in the risk map, from white to purple means that the more purple an area, the more dangerous (in case of a risk map) it is. We compared our risk maps with the static population map resulting from the census data [8], and we computed the correlations between the two.

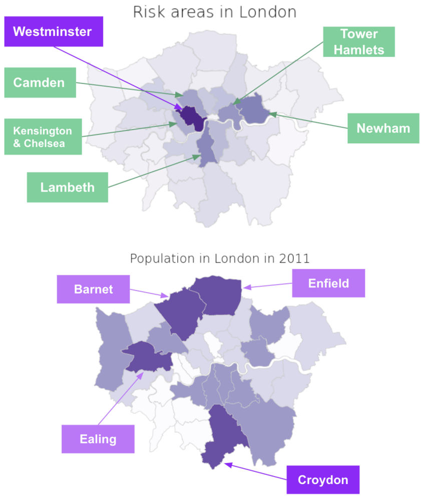

Figure 1: Risk map (up) versus Census map (down), at borough level.

With all the considered granularities, the average risk maps look very different in comparison with their respective analogous population map. This can be seen at first sight in the example (Figure 1).

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

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 these differences numerically, for each granularity level, we also computed the correlations between the daily risk maps and the population map. ***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***. 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.

4.3 Bounding box analysis

We computed for each (detected as) hospitalized individual the bounding box containing the set of locations on which the user moves. 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.

5 CONCLUSIONS

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 at various scales. The proposed method yields a 98.6% agreement with released public records of patients admitted to hospitals in London, within the same time frame. A validation and fine-tuning study is performed to found 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. 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. 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. We also observe that hospitalized individuals tended to have a higher average mobility than non-hospitalized ones. Overall, 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.

ACKNOWLEDGMENTS

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

REFERENCES

[1] J. Álvarez et al. 2021. Estimating Active Cases of COVID-19. *arXiv preprint arXiv:2108.03284.*

[2] L. Bengtsson et al. 2015. Using mobile phone data to predict the spatial spread of cholera. *Scientific reports.*

[3] G. Bobashev et al. 2020. Geospatial forecasting of COVID-19 spread and risk of reaching hospital capacity. *SIGSPATIAL Special.*

[4] L. Bradford et al. 2020. COVID-19 contact tracing apps: a stress test for privacy, the GDPR, and data protection regimes. *Journal of Law and the Biosciences.*

[5] M. A. Carrillo et al. 2021. The use of mobile phones for the prevention and control of arboviral diseases: a scoping review. *BMC public health.*

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

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

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

[9] S. Davalbhakta et al. 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.*

[10] E. Dong et al. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases.*

[11] Z. Fan et al. 2020. Human mobility based individual-level epidemic simulation platform. *SIGSPATIAL Special.*

[12] S. Gao et al. 2020. Mapping county-level mobility pattern changes in the United States in response to COVID-19. *SIGSPATIAL Special.*

[13] S. Hazarie et al. 2021. Interplay between population density and mobility in determining the spread of epidemics in cities. *Communications Physics.*

[14] S. Isaacman et al. 2018. Modeling human migration patterns during drought conditions in La Guajira, Colombia. In *ACM SIGCAS COMPASS.*

[15] J. S. Jia et al. 2020. Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature.*

[16] M. Mokbel et al. 2020. Contact tracing: Beyond the apps. *SIGSPATIAL Special.*

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

[18] O. Ojo et al. 2020. CoronaSurveys: Using Surveys with Indirect Reporting to Estimate the Incidence and Evolution of Epidemics. *arXiv preprint arXiv:2005.12783.*

[19] J. Rajarethinam et al. 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.*

[20] E. M. Rees et al. 2020. COVID-19 length of hospital stay: a systematic review and data synthesis. *BMC medicine.*

[21] Z. Sun et al. 2020. Community venue exposure risk estimator for the COVID-19 pandemic. *Health & Place.*

[22] G. Thakur 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.*