Heydari, Sara; Huang, Zhiren; Hiraoka, Takayuki; Ponce de Leon Chavez, Alejandro; Ala-Nissila, Tapio; Leskelä, Lasse; Kivelä, Mikko; Saramäki, Jari

Estimating inter-regional mobility during disruption: Comparing and combining different data sources

Published in:
Travel Behaviour and Society

DOI:
10.1016/j.tbs.2022.11.005

Published: 01/04/2023

Document Version
Publisher's PDF, also known as Version of record

Travel Behaviour and Society 31 (2023) 93–105

Contents lists available at ScienceDirect

Travel Behaviour and Society

journal homepage: www.elsevier.com/locate/tbs

Estimating inter-regional mobility during disruption: Comparing and combining different data sources

Sara Heydari [a,*], Zhiren Huang [a], Takayuki Hiraoka [a], Alejandro Ponce de León Chávez [b], Tapio Ala-Nissila [c,d], Lasse Leskelä [b], Mikko Kivelä [a], Jari Saramäki [a]

[a] Department of Computer Science, Aalto University, Espoo 00076, Finland
[b] Department of Mathematics and Systems Analysis, Aalto University, Espoo 00076, Finland
[c] Quantum Technology Finland Centre of Excellence and Department of Applied Physics, Aalto University, Espoo 00076, Finland
[d] Interdisciplinary Centre for Mathematical Modelling and Department of Mathematical Sciences, Loughborough University, Loughborough LE11 3TU, United Kingdom

ARTICLE INFO

Keywords:
Human mobility
Travel demand estimation
Origin-destination matrix
Road traffic data
Mobile phone data
Mobility models

ABSTRACT

A quantitative understanding of people's mobility patterns is crucial for many applications. However, it is difficult to accurately estimate mobility, in particular during disruption such as the onset of the COVID-19 pandemic. Here, we investigate the use of multiple sources of data from mobile phones, road traffic sensors, and companies such as Google and Facebook in modelling mobility patterns, with the aim of estimating mobility flows in Finland in early 2020, before and during the disruption induced by the pandemic. We find that the highest accuracy is provided by a model that combines a past baseline from mobile phone data with up-to-date road traffic data, followed by the radiation and gravity models similarly augmented with traffic data. Our results highlight the usefulness of publicly available road traffic data in mobility modelling and, in general, pave the way for a data fusion approach to estimating mobility flows.

1. Introduction

An accurate, quantitative understanding of the mobility patterns of people is crucial for many applications, such as transport engineering (Wang et al., 2012; Ren et al., 2014; Guirao et al., 2018), emergency management (Lu et al., 2012; Huang et al., 2018), and modelling and forecasting the spread of contagious diseases (Balcan et al., 2009; Belik et al., 2011; Riley et al., 2015). While the relevant time scales for engineering applications are fairly long, natural disasters or pandemics typically result in sudden changes in mobility, caused by the event itself and by government interventions (Lu et al., 2012; Tian et al., 2020; Kraemer et al., 2020; Barbieri et al., 2021). As the traditional mobility models are practically static, being based on geography and demographics (Barbosa et al., 2018), this calls for the use of dynamic data on the movement patterns of people. In this paper, we combine data from various sources with the aim of investigating inter-regional mobility in Finland during the onset of the SARS-CoV-2 pandemic in early 2020. To this end, we develop various models, using the traditional static mobility models as starting point for some.

The mobility flows of people have been modelled as early as in the 19th century (Ravenstein, 1885) where the distribution of population and the distances between cities were found to be important factors for determining mobility. Since then, a large number of theoretical models have been proposed, most notably the gravity (Zipf, 1946) and radiation models (Simini et al., 2012; Ren et al., 2014) which have become the de facto standard for estimating intercity travel flows. Both models take geographical distances and the population distribution as inputs and provide static estimates for mobility flows. Generally, they capture the big picture of mobility rather well; however, they are not perfect and are at times prone to underfitting (see, e.g., Masucci et al. (2013)). Both models also require up-to-date census data, which may not always be available. Moreover, when there are sudden changes in mobility due to exogenous factors, these models are obviously of limited use on their own.

Many studies have employed different data sources to improve the accuracy and timeliness of mobility estimates (Vespignani, 2009). Brockmann and Theis (2008) estimated the distribution of travel between cities in the U.S. by using the circulation records of dollar bills from a website (wheresgeorge.com). Another approach is to estimate mobility patterns using the digital traces that people leave behind while moving, e.g. smartcard data (Huang et al., 2018) and Bluetooth data (Laharotte et al., 2015). Furthermore, companies such as Twitter,

* Corresponding author.

https://doi.org/10.1016/j.tbs.2022.11.005
Received 29 June 2022; Received in revised form 12 October 2022; Accepted 12 November 2022
Available online 5 December 2022

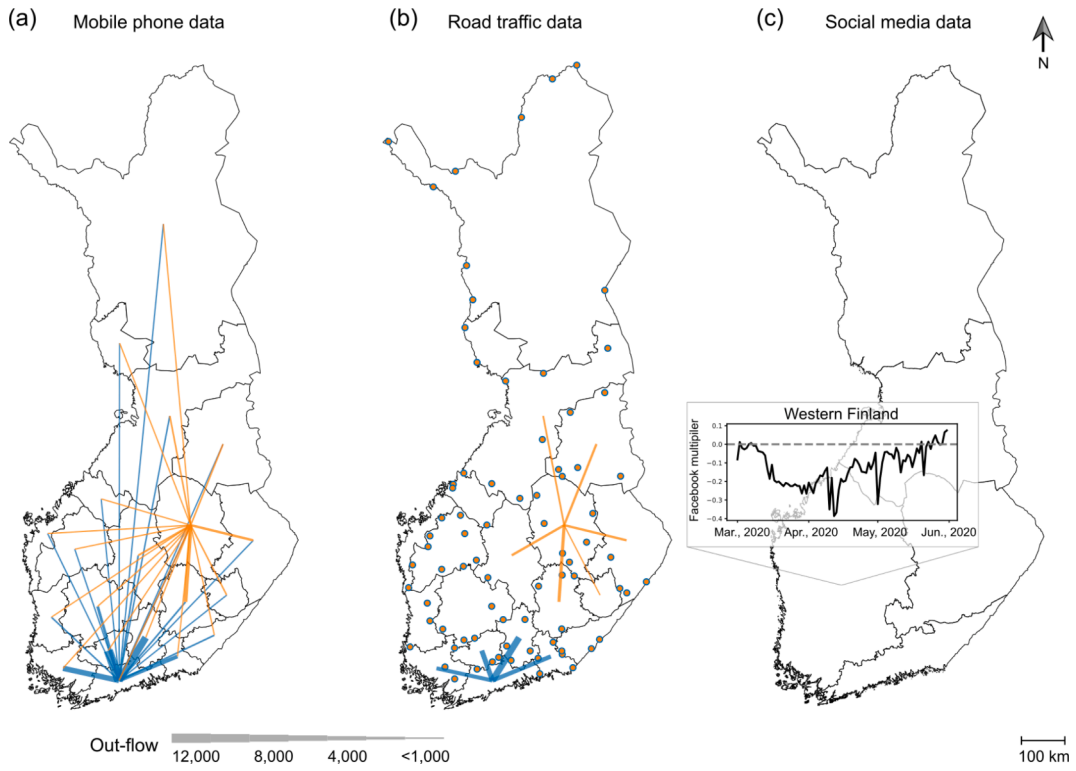

Fig. 1. Illustration of the three different datasets on mobility in Finland used in the present study. The left panel shows mobile phone data overlaid on the 20 hospital districts in mainland Finland. These data provide the numbers of trips between each pair of hospital districts. The orange and blue lines demonstrate trips originating from two hospital districts on a typical day, so that the line thickness is proportional to the number of trips. The road traffic data (middle panel), on the other hand, only report the observed number of motor vehicles travelling between neighbouring districts. These traffic flows are calculated based on the number of passing vehicles as captured by road traffic sensors located close to the borders of the hospital districts, as marked on the map. Similarly to the left panel, the orange and blue lines demonstrate the flows on a typical date. The Facebook data (right panel) report changes in the level of movement of people inside the regions indicated with black lines; there is no information on inter-regional mobility. The regions used by Facebook are larger than the hospital districts. The inset plot illustrates the mobility change index provided with Facebook for one of the regions, from March 2020 to June 2020.

Google, and Facebook have access to the geolocation of their users via GPS sensors on the users' mobile phones or via their IP address. Some studies have used such geolocation data to estimate mobility patterns (Giannotti et al., 2011; Hawelka et al., 2014; Provenzano et al., 2018; Huang et al., 2020; State et al., 2013). While data collected by these companies can have a high temporal and spatial coverage, they may not be representative of the general population and the availability of such data is limited.

For capturing inter-regional mobility, the geolocation records of mobile phone operators provide an excellent source of data. These data are collected by base stations that continuously keep track of the phones within their range. Such data can be aggregated into dynamic Origin–Destination (OD) matrices that contain anonymized information on the number of phones that have moved between any two regions during a given time frame. However, these data are generally not freely accessible for modelling purposes and it is difficult to obtain real-time access to them, which would be required for real-time monitoring of mobility flows.

Real-time monitoring is, however, possible using publicly available data on road traffic. Such data are collected via automated sensors and they are mainly used for the purpose of monitoring traffic conditions and road usage. Contrary to data collected from personal devices, road traffic data do not require anonymization and they are typically freely shared by the traffic authorities. The shortcomings of road traffic data are that they only represent one mode of transportation and are sampled only from specific locations on the major roads.

In this paper, we set out to provide a paradigm for accurately estimating the changing mobility between 20 hospital districts in mainland Finland during the first wave of the COVID-19 pandemic in the spring of 2020, a period in which regular mobility patterns were highly distorted. For this purpose, we use different types of data, including public datasets such as the distribution of population, real-time road traffic data, mobility reports published by Facebook and Google, and proprietary data provided by a major telecommunication company. We explore how these different types of data can be used to develop temporal models, and compare the performance of these models in estimating changes in human mobility under unusual circumstances. For measuring the performance of our models, we use a part of the mobile phone data provided by the telecommunication company as the ground truth. We also compare the performance of our models with those of the widely-used static radiation and gravity mobility models.

Our results point out the utility of road traffic data in estimation of inter-regional mobility. The advantage is particularly prominent when the regular patterns of mobility are distorted, such as during the spring of 2020. Augmenting the traditional static mobility models with dynamic road traffic data improves their performance, and road traffic data works particularly well in combination with pre-pandemic mobile-phone mobility data.

The structure of the paper is as follows. In Section 2 we describe the contents and origin of the different datasets used in this study. Then, in Section 3 we introduce the models developed for estimating mobility flows and discuss the error function for measuring their performance. We then present results from a comparison of the models against the ground truth data in Section 4, and finally discuss the results and present our conclusions in Section 5. The paper is followed by an Appendix in which we elaborate the datasets in more details (Section B) and present the mathematical proof for our modified version of the radiation model (Section C).

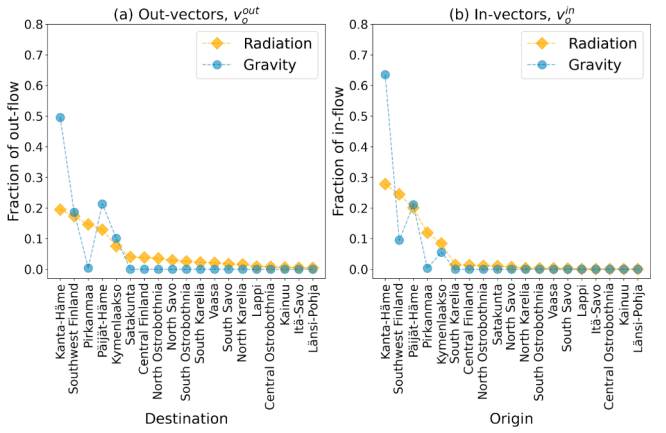

Fig. 2. Out-vector (a) and in-vector (b) of Uusimaa hospital district as estimated by static radiation and gravity models, depicting the fractions of flow to/from other hospital districts. The hospital districts of the horizontal axes have been sorted on the basis of the radiation model estimates, in decreasing order of flow.

2. Datasets

We use five different data sources. Data on the geographical distribution of workers and jobs, from Statistics Finland, are used for estimating mobility flows with the static gravity and radiation models. Dynamic mobility data from road-traffic sensors, Facebook, and Google are used to make these models dynamic. Finally, we use a commercial mobile-phone-based mobility dataset partially for parametrizing our models and partially as the ground truth for their validation. The main features of the mobility datasets are introduced below. For a detailed description, we refer the reader to Appendix B.

The public road traffic data contain the number of vehicles captured by traffic sensors in 5-min intervals. There are in total over 450 of these sensors on the Finnish roads. Data from the year 1995 to the current day are available from the website of Traffic Management Finland (Finland, 2022). We aggregate data provided by those sensors that are located near the borders of the 20 hospital districts and use them as a proxy of the total traffic entering and leaving each district (cf. Fig. 1). Google and Facebook data are available only since March 2020, the time when COVID-19 became widespread. They report how the movement of people inside each region has changed relative to the pre-pandemic situation. It should be noted however, that Google and Facebook data have different geographical resolutions (See Fig. S1 for a comparison of these different geographical resolutions and for how we map these different scales to each other).

The final dataset that we use is based on aggregated and anonymized mobile-phone geolocation records, provided for research purposes by Telia, the 2nd largest mobile phone operator in Finland. These data span two 4-month periods in 2019 and 2020, from the beginning of February to the end of May for both years. We refer to the 2019 data as *pre-pandemic mobile phone data* and the 2020 data as *pandemic mobile phone data*. The data contain the number of trips in six-hour time bins between the 295 Finnish mainland municipalities (see Appendix A, Fig. S1). Unlike the other three mobility datasets, the mobile-phone data provide a full origin–destination (OD) matrix. We use the pandemic mobile phone data as the ground truth for measuring the performance of our various models. From the pre-pandemic mobile phone data, we extract the essential patterns of mobility which can be used instead of the gravity and radiation models. Moreover, we use these data to augment the road traffic data.

3. Methods

Our aim is to develop and investigate models that use different data sources to provide accurate mobility estimates at the inter-regional level. Our particular focus is on dynamic models that perform well even under extraordinary circumstances that result in sudden deviations from typical mobility patterns. To this end, we use the first months of the COVID-19 pandemic to test the performance of the models.

We use our models to estimate origin–destination matrices whose elements describe the number of people moving between pairs of Finnish hospital districts during 12-hour time bins. Hospital districts are a natural choice for the geographical granularity of any mobility estimates intended the be used as input in pandemic modelling, because during the course of COVID-19 pandemic, important statistics such as the number of hospitalisations were reported at this level (Statistics in Finland, 2021). Regarding temporal granularity, since the static gravity and radiation models that we use are models of commuting flows, we choose a matching resolution of 12 hours. Note that we use a higher initial geographical and temporal resolutions for some of the other models, but then aggregate the results into hospital districts level and 12-h time bins. The aggregation enables us to compare the performance of the models.

In this Section, we will first introduce the concept of mobility vectors that are used as a building block of several of our models. We will then introduce each of our models, and finally discuss the error function that is used for comparing the performance of the different models.

3.1. Mobility vectors

The connectivity profile of a region can be quantified using the concept of *mobility vectors*. The mobility vector of a region captures the way how it is embedded in the broader mobility network. We will use mobility vectors in Section 3.3 for combining mobile phone and road traffic data, and in Section 3.4.3 to make radiation and gravity models dynamic with the help of the road traffic data.

Two mobility vector can be defined for each region and for each period of time: the *out-vector* quantifies how the outgoing flow from the region is distributed among the destinations during the time period, and the *in-vector* similarly quantifies the incoming flows to this region.

Given an OD matrix $\mathbf{F}$ with diagonal elements equal to zero (only containing inter-regional flows) and whose element f_{od} denotes flow between origin o and destination d, the out-vector of o is defined as the normalized vector

$$\mathbf{v}_o^{\text{out}} = \frac{1}{\sum_i f_{oi}} \left(f_{o1}, \ldots, f_{oj}, \ldots\right) = \left(v_{o1}^{out}, \ldots, v_{oj}^{out}, \ldots\right), \tag{1}$$

where v_{od}^{out} indicates the fraction of flow to destination d. Similarly, the in-vector of d is

$$\mathbf{v}_d^{\text{in}} = \frac{1}{\sum_i f_{id}} \left(f_{1d}, \ldots, f_{jd}, \ldots\right) = \left(v_{1d}^{in}, \ldots, v_{jd}^{in}, \ldots\right). \tag{2}$$

The geographical granularity and temporal span of the calculated mobility vectors are the same as those of the corresponding OD matrix. For example, the mobility vectors constructed based on estimations of the static radiation or gravity models would be static as well (see Sections 3.4.1 and 3.4.2 for more details on these static models). An example of mobility vectors estimated by static radiation and gravity models for one of the hospital districts in Finland is shown in Fig. 2.

3.2. Using pre-pandemic mobile phone data to estimate the OD matrix

Under normal circumstances, the inter-regional mobility patterns can remain fairly similar over several years. Therefore, as the first approximation, our simplest model approximates mobility flows directly from the patterns of the previous year. This model is not expected to perform well for spring 2020—rather, it serves as a point of comparison with the more involved models that utilize data concurrent with the COVID-19-driven mobility changes.

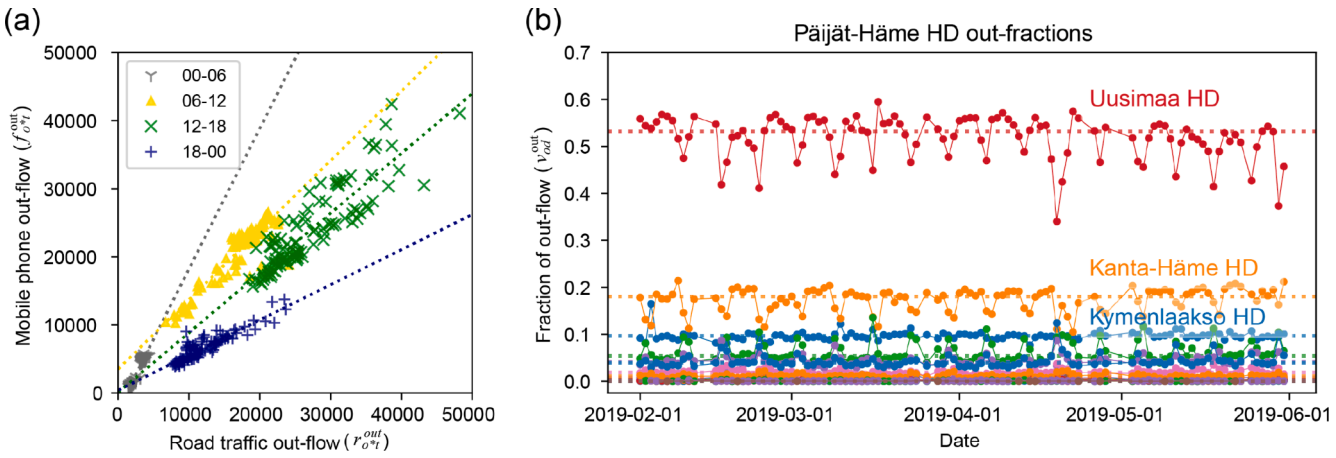

Fig. 3. The components of the model explained in Section 3.3. The model combines pre-pandemic mobile phone data with road traffic data to get temporal estimates of inter-regional mobility. In the figure, we visualise the steps for estimating the flow originating from Päijät-Häme hospital district as an example. **Panel (a)** Total out-flow from Päijät-Häme in 4-h time bins in the pre-pandemic mobile phone data (f^{out}_{i*t}) as a function of the road traffic flow (r^{out}_{i*t}) in the same time bins. The data points are grouped based on their temporal features, in this case the time of the day, and then the line of best fit is found for each group (the dotted lines). **Panel (b)** Fractions of out-flow of Päijät-Häme to each destination (solid lines) and the corresponding median flows (dotted lines). Here, only the fractions for the morning time bins and their corresponding medians are visualized. Information in panels (a) and (b) is combined to estimate inter-regional mobility. For details see Eqs. (4)–(8) and Section 3.3.1.

The mobile phone data originally come with a 6-h granularity, pre-binned for each date into 4 bins: night (00-06), morning (06-12), afternoon (12-18), and evening (18-00).

When estimating mobility flows from the pre-pandemic mobile-phone data, our aim is to retain the periodic temporal patterns typical for mobility, such as periodic daily variations, while removing random temporal fluctuations. Therefore, for an origin–destination pair (o, d) and a given time bin t, we estimate the flow $\widetilde{f}^{\text{mobile}}_{odt}$ as the median of the flows from o to d in all the time bins in the pre-pandemic data that correspond to the same weekday and a similar 6-h bin:

$$\widetilde{f}^{\text{mobile}}_{odt} = \text{med}[f_{odi}], \forall i: \quad \begin{aligned} &\text{WD}(i) = \text{WD}(t) \& \\ &\text{TD}(i) = \text{TD}(t), \end{aligned} \tag{3}$$

where WD(i) indicates the day of the week of bin i and TD(i) the time of the day of i (night, morning, afternoon, evening). Note that the estimate is therefore the same for, e.g., any Monday morning.

3.3. Augmenting pre-pandemic mobile phone data with road traffic data

Next, we continue with pre-pandemic mobile phone data, using it as a baseline to be augmented with dynamic data from the road traffic sensors in order to obtain estimates that work for the 2020 pandemic period. We start by investigating the relation between the two sets of data during the spring of 2019, developing a model that estimates mobility flows for any given date from continuously available, up-to-date traffic data and an earlier mobile-phone baseline. As discussed in Section 2, road traffic data only capture partial traffic flows between neighbouring hospital districts, and cannot therefore alone provide a full OD matrix. Therefore we use pre-pandemic mobile data as part of the model.

The procedure of combining road traffic data with mobile phone data consists of two steps. First, we find the scaling from the number of vehicles observed on the roads to the total in-flow and out-flow of each region in the pre-pandemic mobile data. Second, we use mobility vectors together with these flows to get the full OD matrix. This is done, for each region, by distributing the total out-flow (in-flow) among destinations (origins) proportionally to the corresponding mobility vector elements (see Section 3.1 for the definition of mobility vectors).

To find the relation between the observed number of vehicles and the number of travellers in the pre-pandemic mobile-phone OD matrix, we start by aggregating both types of data into the same regional and temporal level. We set the temporal resolution to 6 h (the resolution of the mobile phone data). For the spatial resolution, we use the 20 hospital districts (cf. Fig. S1 in the Appendix).

To find the transformation between the road traffic and mobile phone data, we first use the mobile phone data to compute the total number of individuals leaving and entering region i at time t, f^{out}_{i*t} and f^{in}_{*it}. Similarly, we use the road traffic data to calculate the total numbers of vehicles entering and leaving the region, r^{out}_{i*t} and r^{in}_{*it}.

We then investigate how the number of travellers from the mobile phone data is related to the number of vehicles from the road traffic data, separately for each region i and direction (out-flow and in-fow). The dependence can be expected to be approximately linear (the average number of passengers per car should not strongly depend on traffic volume), which is confirmed by visualizing the data (see Fig. 3). Therefore, we perform simple linear regression with r^{out}_{i*t} and r^{in}_{*it} as the independent variables and f^{out}_{i*t} and f^{in}_{*it} as the dependent variables, respectively. With the resulting slopes (m^{out}_i and m^{in}_i) and intercepts (b^{out}_i and b^{in}_i), we can estimate the total mobility in-flows and out-flows for each region i at each time t as

$$\widetilde{f}^{\text{out}}_{i*t} = m^{\text{out}}_i r^{\text{out}}_{i*t} + b^{\text{out}}_i, \tag{4}$$

and

$$\widetilde{f}^{\text{in}}_{*it} = m^{\text{in}}_i r^{\text{in}}_{*it} + b^{\text{in}}_i. \tag{5}$$

Each region thus has its own linear model for the in- and outflows.

Next, we want to determine how the outflow (inflow) of a given district is distributed among destinations (origins). To this end, we use the mobility out-vector and in-vector of each region, calculated using the pre-pandemic data (see Section 3.1). The out-vector of origin o at time t is an array of fractions v^{out}_{odt} of flow from the origin o to each destination d. Similarly, we denote the fractions of in-flow by v^{in}_{odt}. For each region o, we calculate the median values of the fractions of in-flow and out-flow over all time bins in the pre-pandemic data. We denote these median fractions as

$$\widetilde{v}^{\text{out}}_{od} = \text{med}\left[v^{\text{out}}_{odt}\right]; \tag{6}$$

$$\widetilde{v}^{\text{in}}_{od} = \text{med}\left[v^{\text{in}}_{odt}\right]. \tag{7}$$

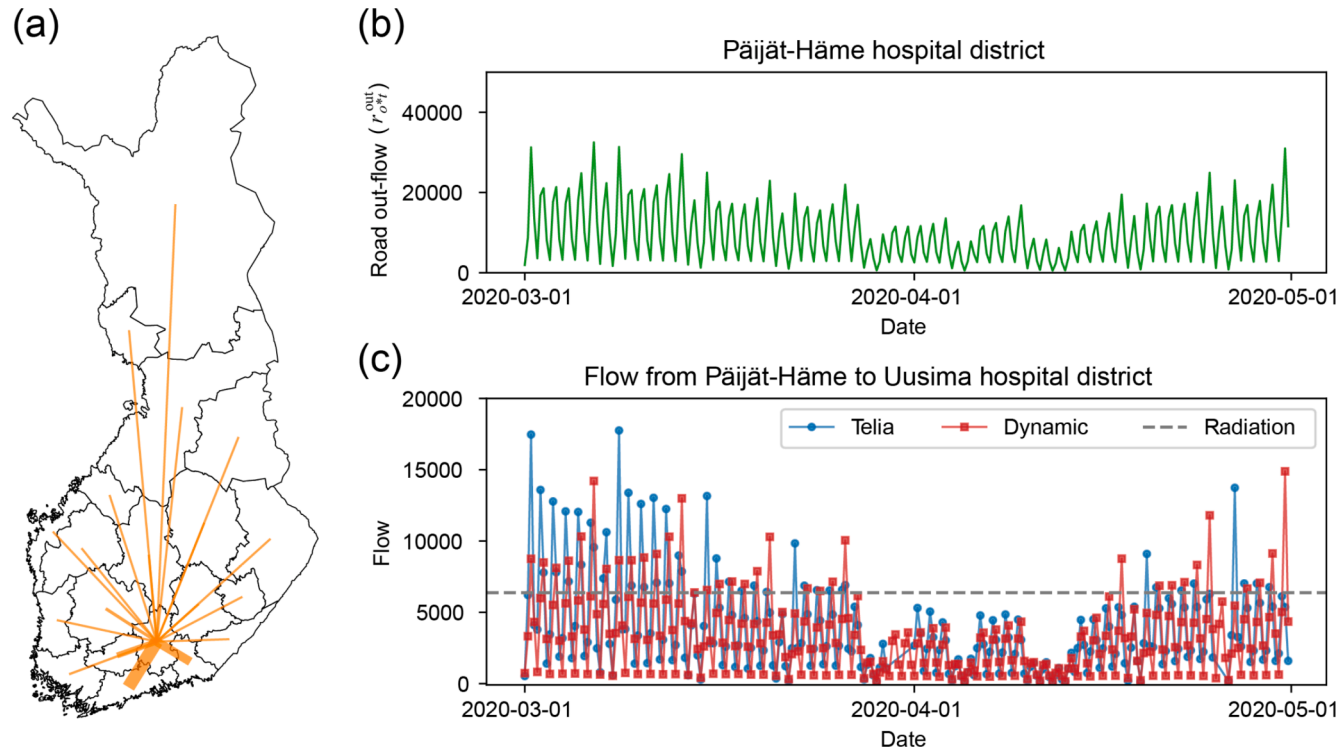

Fig. 4. An illustration of how we augment the static radiation model (RM) and gravity model (GM) with the road traffic data to get dynamic estimates. Panel (a) shows a schematic picture of the out-vector of the Päijät-Häme hospital district based on the RM. Line thickness corresponds to the magnitude of the out-vector component. We combine these fractions with the temporal out-flow from Päijät-Häme, using road traffic data (displayed in panel (b)), to get dynamic estimates. In panel (c), the resulting dynamic estimate of the flow between Päijät-Häme and Uusimaa is shown in red. The blue lines indicate the ground truth from the mobile phone data. The temporal variations are in good agreement. The static RM estimate is shown by the dashed grey line.

Then, these values are renormalized so that the "median" mobility vector is also a unit vector, i.e., $\sum_d \tilde{v}^{out}_{od} = 1$ and $\sum_d \tilde{v}^{in}_{od} = 1$. Now, by combining the models for the individual regions and the median mobility vectors, we can estimate the entries of the OD matrix. To this end, we take the total out- and in-flows of Eqs. (4) and (5) and distribute them among destinations/origins by multiplying them with the values provided by Eqs. (6) and (7). As we get two estimations for each pair (o, d)—one for the out-flow from o to d and another for the in-flow to d from o, we finally average over them to arrive at the estimated flow

$$\tilde{f}_{odt} = \left(\tilde{f}^{out}_{o*t}\tilde{v}^{out}_{od} + \tilde{f}^{in}_{*dt}\tilde{v}^{in}_{od}\right)\Big/2. \tag{8}$$

Please note that unlike in the theoretical radiation and gravity models, where the in-flow from o to d is equal to the out-flow from d to o, this is not necessarily the case for this data-driven model—travelers from o to d do not necessarily return on the same day.

3.3.1. Taking weekly and daily periodicities into account

In the previous section, we constructed a model which estimates the OD matrix for time bin t based on the number of observed vehicles leaving and entering regions in that time bin. In training this model, we used all the datapoints in the pre-pandemic mobile phone data. However, it is well known that mobility is influenced by weekly and daily periodicities—e.g., there is less traffic on weekends and the morning rush hour is different from the early afternoon. As the average number of passengers per vehicle can also be expected to depend on these periodicities, we should arrive at a better estimate if this is taken into account in the regression of Eqs. (5) and (4). Similarly, the mobility vectors can be expected to differ, e.g., between weekends and weekdays (commuting vs leisure travel), which can be accounted for in Eqs. (6) and (7).

We use three time-dependent models: one where the time bins are grouped by weekday, one where the grouping is by time of day, and one where both are used. We then perform both the linear regression from traffic counts to mobility as well as the median estimation from pre-pandemic data separately for different groups of time bins. For the regression, the coefficients of Eqs. (5) and (4) therefore become time-dependent: for each group of time bins, we get separate values of m^{out}_{it} and b^{out}_{it}. Similarly, the median mobility vectors of Eq. (6) and (7) will explicitly depend on time and we have $\tilde{v}^{out}_{odt}$ and $\tilde{v}^{in}_{odt}$ where the medians are now taken over the data points in similarly defined groups of time bins. The final flow estimate is then calculated as an average similarly to Eq. (8). An example of the procedure is shown in Fig. 4.

3.4. Models based on publicly available data

Mobility data based on mobile-phone location is considered as one of the best proxies for the movement of people. However, mobile phone data are usually not public. In this section, we propose mobility models developed based on public data such as distribution of population and jobs, as well as road traffic data and the public mobility trends published by Google and Facebook as a response to the COVID-19 pandemic.

We first introduce versions of the static radiation and gravity models which predict an OD matrix based on the geographical distribution of workers and jobs. Then we augment these models in turns with road traffic data and mobility trends from Facebook and Google to make the estimates dynamic.

3.4.1. Static OD matrix from the radiation model

The radiation model (RM) is an intervening opportunity model where a given job-seeker chooses a workplace considering a trade-off between the job's benefits and the commute length. In the basic model proposed by Simini et al. (2012), a job-seeker chooses the closest job to the home region which is better than the best job available in the home region.

In Simini et al. (2012), it is assumed that the number of employment opportunities in each region is proportional to the local population. However, having access to the spatial distribution of jobs and commuters in Finland (Section B.2), we propose a modified version of the RM which uses these data instead of the general population distribution. For this model, we initially use a higher spatial resolution (at the level of municipalities) and then aggregate the results to the level of hospital districts. The details of the model are explained in Appendix C.

According to the model, the conditional probability that a worker living in municipality i works in municipality j is:

$$p_{ij} = \begin{cases} 1 - \frac{n_{i+}}{n_i}, & j = i; \\ c_i \frac{m_i m_j}{(m_i + s_{ij})(m_i + m_j + s_{ij})}, & j \neq i, \end{cases} \tag{9}$$

where n_i denotes number of workers residing in i out of which n_{i+} are out-commuters, m_i denotes number of jobs located in $i, s_{ij} = \sum_{k \neq i: d_{ik} < d_{ij}} m_k$ denotes the number of jobs which are located closer to i than j is, and the normalization constant is given by $c_i = \frac{n_{i+}}{n_i}\left(1 - \frac{m_i}{\sum_i m_i}\right)^{-1}$. If all n_i workers living in i choose their work destination using (9), then the expected number of people living in i and working in j equals

$$n_{ij} = n_i p_{ij}.$$

These numbers are then aggregated to the level of hospital districts. We denote the aggregated flow from hospital district i to hospital district j by η_{ij}.

The RM is a model of commuting, and therefore a natural choice is to divide each day into two twelve-hour time bins (00–12 and 12–24) and assume that the flow observed from midnight to noon consists of people commuting to work and the flow from noon to midnight of returning commuters. Then, the flow of people from hospital district i to hospital district j at time bin t is estimated as

$$\tilde{f}^{\mathrm{RM}}_{ijt} = \begin{cases} \eta_{ij}, & t \in [00-12]; \\ \eta_{ji}, & \mathrm{t} \in [12-24]. \end{cases} \tag{10}$$

Later in Section 3.4.3 we combine the estimates of the static RM with road traffic data for a dynamic version of the RM.

3.4.2. Static OD matrix from the gravity model

The gravity model (GM) refers to a class of models for describing spatial interactions between regions and has many variations (Zipf, 1946; Wilson, 1971). The most common model of movement between regions is the production-constrained gravity model (Wilson, 1971; Lenormand et al., 2016) where the share of between-municipality commuters who live in municipality i and work in municipality j is given by the proportionality

$$p_{ij} \propto n_i m_j \exp(-\beta d_{ij}), \quad j \neq i. \tag{11}$$

Here n_i is the number of workers who live in municipality i, m_j is the number of jobs in municipality j, d_{ij} is the distance between the two municipalities, and β is the parameter in the exponential function (data on the distribution of jobs and workers are discussed in Section B.2). According to Lenormand et al. (2016), β can be directly inferred from the average spatial unit surface as

$$\beta = 0.3\langle S \rangle^{-0.18}, \tag{12}$$

where $\langle S \rangle$ is the average surface area of the municipalities which equals 1081.016 km^2 in our case. We use the value of the exponent, -0.18, that was estimated in Lenormand et al. (2012) using data for 4 countries.

From official statistics we can get n_{i+}, the total number of out-commuters of each municipality (see Section B.2 for details on the data). We determine the expected number of commuters n_{ij} who live in municipality i and work in municipality j by distributing out-commuters according to the probabilities in Eq. (11):

$$n_{ij} = \frac{p_{ij}}{\sum_k p_{ik}} n_{i+}. \tag{13}$$

As the last step, we aggregate these estimates to get the OD matrix for the mobility between the hospital districts. We denote the aggregated flow from hospital district i to hospital district j by η_{ij}. As for the temporal resolution, similarly to the case of the radiation model we assume that the population commutes to work during the twelve-hour time bin from midnight to noon and commutes back home from noon to midnight. Thus, the flow from hospital district i to hospital district j at time bin t according to the static GM is

$$\tilde{f}^{\mathrm{RM}}_{ijt} = \begin{cases} \eta_{ij}, & t \in [00-12]; \\ \eta_{ji}, & \mathrm{t} \in [12-24]. \end{cases} \tag{14}$$

3.4.3. Dynamic radiation and gravity models with road traffic data

Next, we want to increase the accuracy of the RM and GM models by making them dynamic with the help of the road traffic data. We begin by using the static RM and GM models to calculate the mobility vectors for each region similarly to how the pre-pandemic mobile data is used in Section 3.3, but without any time dependence. As the outcome, for each region o, we get the out-vector $\mathbf{v}^{\mathrm{out}}_o$ which consists of fractions of out-flow to each destination v^{out}_{od} (see Eq. 1). Similarly, for each region d, we get the in-vector $\mathbf{v}^{\mathrm{in}}_d$ which consists of in-flow fraction from different origins v^{in}_{od} (see Eq. 2). An example of the out-vectors and in-vectors calculated based on the RM and GM is shown in Fig. 2.

We then use road traffic data to compute the total outgoing and incoming number of vehicles for each hospital district in each time bin, r^{out}_{i*t} and r^{in}_{*it}, respectively. Following the reasoning of Section 3.4.1, we use a time resolution of 12 h. We assume that all vehicles leaving a region in the morning (from midnight to noon) are out-commuters who live in that region. Similarly, all traffic observed in the evening (from noon to midnight) are assumed to be out-commuters returning to their regions of residence. Furthermore, as the models introduced in this section are to be developed using publicly available data only, we cannot perform regression against pre-pandemic mobile phone data as in Section 3.3. Therefore, we directly use the numbers of vehicles as mobility estimates, in effect assuming that each vehicle contains one commuter: $\tilde{f}^{\mathrm{in}}_{*it} = r^{\mathrm{in}}_{*it}$ and $\tilde{f}^{\mathrm{out}}_{i*t} = r^{\mathrm{out}}_{i*t}$.

Following these assumptions, we can estimate the entries of the OD matrix at time t from the observed road out-flow and in-flow in that time bin combined with the mobility vectors calculated from the static radiation or gravity model (for a schematic overview, see Fig. 4). Combining all steps, the estimate of the dynamic radiation model is

$$\tilde{f}^{\mathrm{RM+R}}_{ijt} = \begin{cases} \left(r^{\mathrm{out}}_{i*t} v^{\mathrm{out,RM}}_{ij} + r^{\mathrm{in}}_{*jt} v^{\mathrm{in,RM}}_{ij}\right)/2, & t \in [00-12]; \\ \left(\mathrm{r}^{\mathrm{out}}_{\mathrm{j*t}} \mathrm{v}^{\mathrm{in,RM}}_{\mathrm{ij}} + \mathrm{r}^{\mathrm{in}}_{\mathrm{*it}} \mathrm{v}^{\mathrm{out,RM}}_{\mathrm{ij}}\right)/2, & \mathrm{t} \in [12-24], \end{cases} \tag{15}$$

and similarly, the estimate of the dynamic gravity model is

$$\tilde{f}^{\mathrm{GM+R}}_{ijt} = \begin{cases} \left(r^{\mathrm{out}}_{i*t} v^{\mathrm{out,GM}}_{ij} + r^{\mathrm{in}}_{*jt} v^{\mathrm{in,RM}}_{ij}\right)/2, & t \in [12-24]; \\ \left(\mathrm{r}^{\mathrm{out}}_{\mathrm{j*t}} \mathrm{v}^{\mathrm{in,GM}}_{\mathrm{ij}} + \mathrm{r}^{\mathrm{in}}_{\mathrm{*it}} \mathrm{v}^{\mathrm{out,GM}}_{\mathrm{ij}}\right)/2, & \mathrm{t} \in [12-24]. \end{cases} \tag{16}$$

.

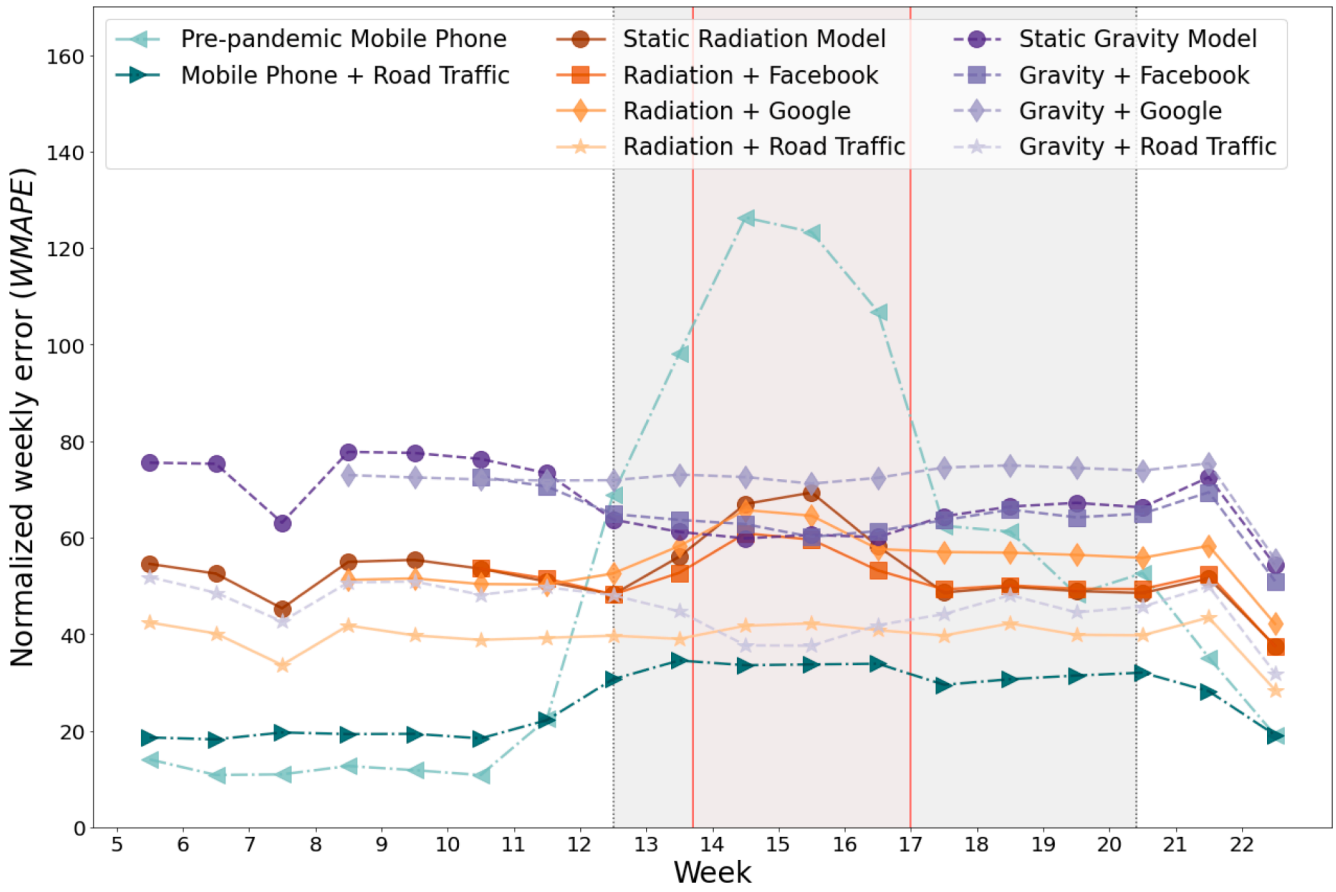

Fig. 5. Comparison of the weekly country-wide weighted mean absolute percentage error WMAPE (see Eq. (19)). For the different models, for weeks 5 to 22 in the year 2020. The shaded grey area between weeks 12 and 20 marks the period of governmental restrictions including the closure of schools. The light red shading marks the closure of the borders of the capital Uusimaa region to non-essential traffic. Overall, the best performance is produced by the model that combines pre-pandemic mobile phone data with road traffic data from 2020 when taking the weekly patterns into account. For the weeks 13–16, this is almost matched by the static models augmented with road traffic data.

3.4.4. *Dynamic radiation and gravity models with Google and Facebook data*

As alternative public data sources for making the radiation and gravity models dynamic, we use data from Google and Facebook. These data provide daily indices of how the level of mobility inside regions has changed. The regions used by Facebook and Google are fairly large, consisting of several municipalities. Therefore, we use the mobility change index of each region for all its constituent municipalities, in order to arrive at the municipality-level resolution of the static RM and GM. The public availability of Google and Facebook data is limited to the pandemic period, as the companies started releasing these data in response to the pandemic.

As the dynamic Google and Facebook mobility indices inform about changes in mobility with respect to the pre-pandemic baseline, we use them as dynamic multipliers to adjust the flows given by the static RM and GM. For this, we assume that mobility between regions has changed similarly to mobility within regions, which is what the indices measure. As an example, assuming that the estimates from static RM are valid for pre-pandemic times, we use the Facebook multipliers to adjust these estimates for pandemic times as

$$\widehat{n}_{ijt}^{\mathrm{FB+RM}} = \begin{cases} n_{ij}^{\mathrm{RM}}\left(c_{it}^{\mathrm{FB}} + c_{jt}^{\mathrm{FB}}\right)\Big/2, & t \in [00-12]; \\ \mathrm{n}_{\mathrm{ji}}^{\mathrm{RM}}\left(\mathrm{c}_{\mathrm{it}}^{\mathrm{FB}} + \mathrm{c}_{\mathrm{jt}}^{\mathrm{FB}}\right)\Big/2, & \mathrm{t} \in [12-24], \end{cases} \tag{17}$$

where t specifies a unique 12-h time bin and c_{it}^{FB} is the mobility change index of municipality i according to Facebook data on the date of time bin t. After calculating the OD matrix at the level of municipalities, similarly to the other models, we aggregate the results to the level of flows between hospital districts.

3.5. *Estimation error*

To measure the performance of our models, we will compare their estimates with the ground truth from the 2020 mobile phone data. We denote the ground truth OD matrix at time t by $\mathbf{F}_t$ and the estimated OD matrix by $\widetilde{\mathbf{F}}_t$. We use the absolute difference to measure the error for each data point, so that the error of the entry ij of the OD matrix equals $\left|\widetilde{f}_{ijt} - f_{ijt}\right|$. The total weekly error in the estimation of inter-regional mobility in the whole country for week w is

$$\epsilon_w = \sum_{t \in w} \sum_{\substack{ij \\ i \neq j}} \left|\widetilde{f}_{ijt} - f_{ijt}\right|. \tag{18}$$

We then calculate the weighted mean absolute percentage error (WMAPE) at the country-wide level and during the span of a week by dividing the weekly error by the total weekly inter-regional mobility:

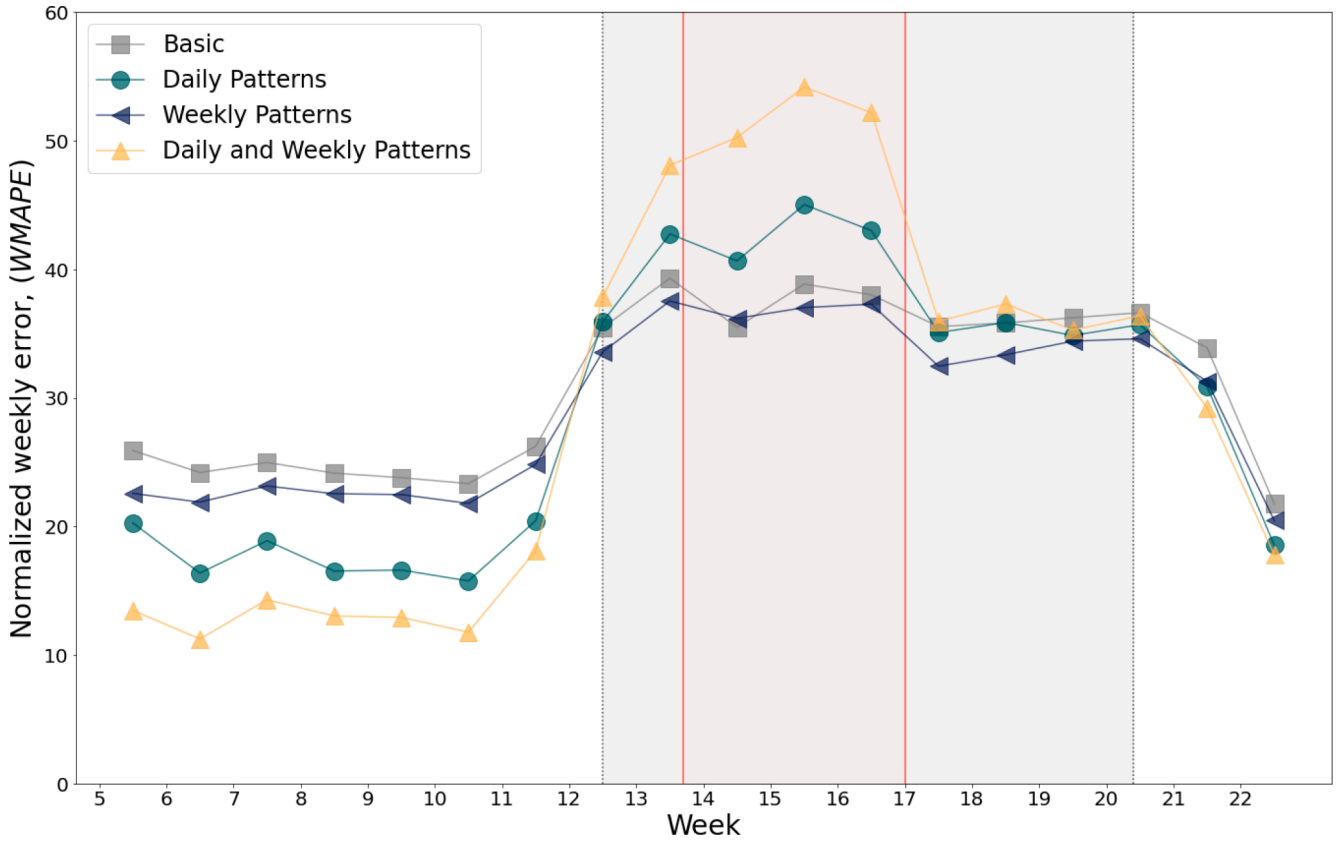

Fig. 6. Comparison between the weekly country-wide weighted mean absolute percentage error WMAPE. for different variations of the model that combines pre-pandemic mobile phone data with road traffic data. The plot shows the WMAPE (see Eq. (19)) in weeks 5 to 22 of the year 2020. The shaded areas are the same as in Fig. 5. The error is smaller for all models that take periodic patterns into account until week 11. After this, the situation changes as the mobility patterns are disrupted by COVID-19: the model tuned with the pre-pandemic daily patterns performs poorly during the pandemic while tuning only to the pre-pandemic weekly patterns decreases the prediction error slightly. This result suggests that daily patterns have changed more significantly than weekly patterns. See Section 3.3.1 for more details.

$$\mathrm{WMAPE}_w = \frac{\sum_{t\in w}\sum_{\substack{ij\\ i\neq j}} \left|\widetilde{f}_{ijt} - f_{ijt}\right|}{\sum_{t\in w}\sum_{\substack{ij\\ i\neq j}} f_{ijt}} \times 100\%. \tag{19}$$

WMAPE has been used in the literature for measuring traffic flow prediction error Do et al. (2019) and as compared to the closely related measure of mean absolute percentage error, it overcomes the problem of diverging error when it comes to the prediction of small values.

4. Results

4.1. Background: the beginning of the pandemic in Finland

We examine the performance of all models from the beginning of February to the end of May 2020. As explained in Section 2, we use mobile phone data from this time period as the ground truth. On week 5 of the year 2020, the first confirmed case of COVID-19 was detected in Finland (YLE, 2021b). This was an isolated case, but the case numbers started rising on week 10 (YLE, 2021a). On week 12, the Finnish government implemented the first restrictions, including the closure of schools, a ban on gatherings of more than ten people, the closure of recreational and sports facilities, and the recommendation to work remotely (YLE, 2020a; YLE, 2020b). These restrictions were followed in the first half of April (from the mid-week of 13 until end of week 16) by a decree on banning unnecessary road traffic between the capital region of Uusimaa, where the prevalence of the disease was the highest, and the rest of the country (YLE, 2020c).

4.2. Performance of models based on pre-pandemic mobile phone data

Let us first consider the scenario where we have access to mobile-phone mobility data from 2019 (pre-pandemic mobile phone data) and we want to predict the mobility in 2020 around the time of the onset of the pandemic. There are two possible strategies: we can estimate the OD matrix directly using pre-pandemic mobile phone data (see Section 3.2), or augment the pre-pandemic mobile phone data with road traffic data (see Section 3.3).

4.2.1. Pre-pandemic mobile phone data as the only input

When the pre-pandemic mobile phone data are used as described in Section 3.2, so that the estimated flows for each weekday and time of day equal the median flows over those weekdays and times of day in 2019, the model performs well for the early weeks of 2020 (see Fig. 5). In other words, under normal circumstances, the mobility patterns can be predicted rather well from an earlier baseline because there is little change. However, from week 12 on, as the pandemic and the related government restrictions induce changes in mobility, the error of estimation grows quickly and clearly exceeds the other models from weeks 13 to 17.

4.2.2. The effect of adding road traffic data

For augmenting the pre-pandemic-data-only model with road traffic data, we first need to choose the best way of taking daily and weekly periodicities into account (see Section 3.3.1). We therefore test four models: the basic version does not account for periodicity, the second considers daily periodicity, the third the differences between weekdays,

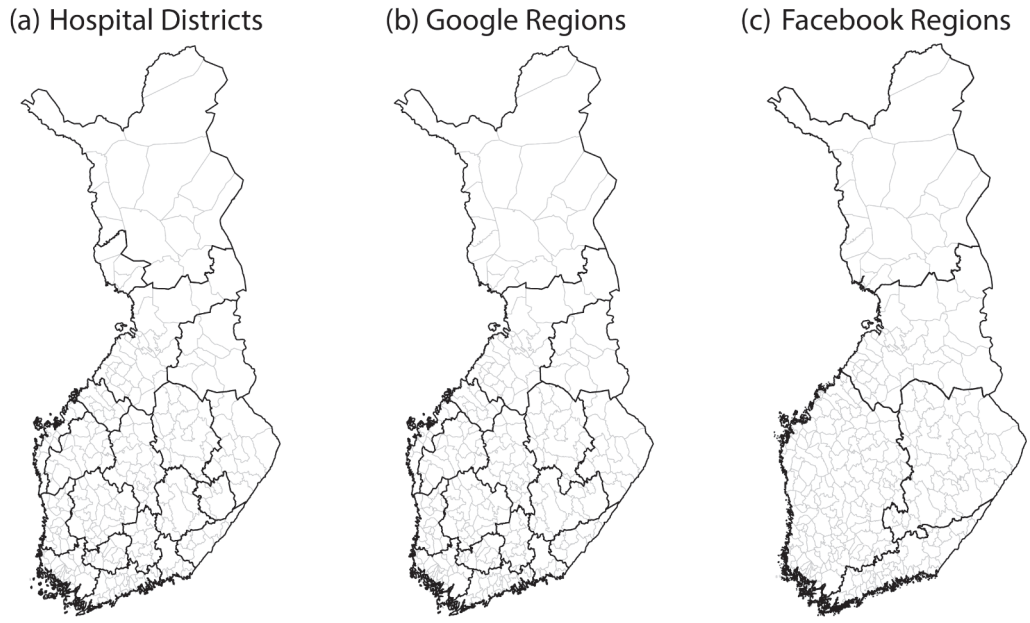

Fig. S1. The map of Finland indicating different geographical regions considered in this work. The 295 municipalities in mainland Finland are marked with light gray lines, excluding the island of Åland, and municipalities are the finest geographical regions among all our datasets. The OD matrix from mobile phone data reports the numbers of trips between these municipalities. The estimations of the static radiation and gravity models are also calculated for inter-municipality mobility flows. These data are finally aggregated to the level of 20 hospital districts as indicated by the map on the left. Google and Facebook data come with their own geographical divisions (the maps in the middle and to the right) that are larger than hospital districts.

Table S1
Sample rows of data provided by Statistics Finland on the distribution of the number of workers and jobs in 2017 for the 295 municipalities in mainland Finland.

i	Workers residing in i (n_i)	Jobs in i (m_i)	Out-commuters (n_{i+})
Helsinki	309 685	397 346	70 820
Espoo	131 153	120 676	67 308
Vantaa	106 531	116 320	57 801
Kirkkonummi	18 086	10 863	12 309

and the fourth both. The resulting errors are shown in Fig. 6.

We observe that considering the periodic temporal patterns, in particular the combination of daily and weekly patterns, reduces the error until week 11. However, the situation changes from around week 12, with the daily and daily + weekly models performing the worst and the model focusing on weekly patterns performing marginally better than the basic model. As the weekly-patterns model works best during the COVID-induced mobility disruption and as its average error during the whole timeline is the smallest of the four models, we choose it for the final comparison.

The result is that the chosen model clearly achieves the best performance for the pandemic period, while being almost on par with the pre-pandemic-only model for early 2020, as seen in Fig. 5. Therefore, considering overall performance, this model clearly wins the comparison.

4.3. Performance of models that rely on public data

Let us next consider the scenario where there is no pre-pandemic mobile phone data, but we need to rely on publicly available data sources. We first estimate static mobility patterns with the radiation and gravity models and then use road traffic data, Google data, and Facebook data to augment these models.

4.3.1. Static radiation and gravity models

The performance of the static radiation and gravity models, constructed following Sections 3.4.1 and 3.4.2, is shown in Fig. 5. Being static, both models yield the same mobility flows for each week, and therefore changes in their weekly errors are purely due to changes in the ground truth. It can be seen that before the pandemic, in February 2020, both models perform considerably worse than the mobile-phone-data based models, with the radiation model having a smaller error than the more traditional gravity model. However, their performance becomes comparable during weeks 13–17.

4.3.2. The effect of adding road traffic data

We find that for the radiation model, using the road traffic data decreases the weekly estimation error on average by 15%. For the weeks 14 and 15, the decrease is more than 30%, and the road-traffic-augmented radiation model performs almost as well as the best model (pre-pandemic mobile data + road traffic), as seen in Fig. 5. Augmenting the static gravity model with the road traffic data also reduces the weekly error, on average by 32%, and its performance in the pandemic weeks 14 to 17 is comparable to the road-traffic-augmented radiation model.

4.3.3. The effect of adding Facebook and Google mobility data

Mobility reports published by Facebook and Google as a response to the COVID-19 pandemic are another alternative for augmenting static mobility models (see Section 3.4.4). The weekly errors of the radiation and gravity models augmented with these data are presented in Fig. 5. Comparing the overall errors reveals that augmenting the static models with these types of data does not improve their performance. On the contrary, for the gravity model, both data sets decrease the performance from week 13. For the radiation model, Facebook data slightly improve the estimates at the pandemic's peak, but this effect is reversed elsewhere and is not persistent.

5. Summary and discussion

Accurate and real-time estimation of country-level mobility patterns is a challenging but crucial task. Such estimates are particularly important in times of crisis, for example when attempting to forecast the spread of contagious disease as well as for evaluating the effect of restrictions imposed to control their spread.

In this study, we have used models based on various types of mobility data as well as traditional static models for the timely estimation of the OD matrix in Finland during the onset of the COVID-19 pandemic. To understand the usefulness of different types of data, we focused on models that are transparent and fairly straightforward. We observed that the best strategy is to extrapolate OD matrices from past data, in our case from a mobile telephone operator, and refine them with dynamic, up-to-date road traffic information. Moreover, using road traffic data in this way also improves the traditional static mobility models substantially.

With the road traffic data and pre-pandemic mobile-phone data, we examined how the different types of periodicity affect the accuracy of the estimates. We found that before COVID-19 became prevalent in Finland, the best estimates were obtained when considering both weekly and daily patterns, that is, variations in mobility between different weekdays and different times of day, as seen in Fig. 6. However, the situation changed dramatically with the onset of the first wave of the pandemic in March 2020. While the performance of all road-traffic-augmented models became substantially worse, the order of the models changed, with the pre-pandemic winner with both daily and weekly patterns now having the worst performance and the weekly model the best. This indicates that in addition to changes in mobility levels, also studied elsewhere (see, e.g., Schlosser et al. (2020, 2022)), there were considerable changes in the temporal patterns of mobility on different timescales. These would be an interesting topic for future studies.

Besides using road traffic data, as an alternative way of augmenting the static mobility models, we investigated using the mobility indices provided by Google and Facebook. In combination with our models, these data sets did not produce further enhancements. There are several likely reasons for this. First, the mobility indices by Facebook and Google quantify changes in mobility trends *inside* regions, while road traffic data provide a proxy of the in-flows and out-flows. Second, Facebook and Google indices are unitless indices of changes in mobility with respect to the pre-pandemic situation, whereas road traffic data yield absolute numbers of vehicles. Third, Facebook and Google indices are provided for relatively coarse geographical regions (see Fig. S1). However, we should emphasize that even though Facebook and Google data are not useful for our models, they might work better for other types of models or other geographical regions.

The performance of any model of mobility is bounded by the accuracy of the data used for the calibration of the model and the ground truth that it is validated against. Therefore, the quality of these data is crucial. In our study, we used mobility data provided by a teleoperator both for the calibration and as the ground truth. These data were scaled at the source (Telia) to account for the operator's market share in different user segments, so that the mobility patterns are representative of the whole population. Naturally, combining mobility data from all (major) teleoperators would be even more optimal, but this can be rarely achieved in practice. Another factor is how recent the training and ground truth data are. Our dynamic models are capable of nowcasting mobility flows by combining the past ground truth with observed signals in the current day. Evidently, long-term changes in mobility patterns degrade the quality of ground truth data from the past, but the quantitative understanding of the relevant timescales would require data from a longer period of time than we had at our disposal.

Essentially, our mobility estimation problem can be seen as a data fusion problem—how to combine various types of data and various models for accurate estimates of mobility? Similar problems have been addressed with different types of data (smart cards, GPS) for the smaller scale of cities in Zhang et al. (2014) and Huang et al. (2018)). Generally, for mobility estimates on the scale of a country as investigated in this paper, there are many possible ways forward. As stated above, we have opted to focus on models that are fairly straightforward and transparent in order to learn the fundamentals of combining data from such a diversity of sources. However, the next steps forward could include the use of machine-learning approaches (Xie et al., 2020; Luca et al., 2021), such as using the deep learning structure ResNet (Zhang et al., 2017) or considering land-use information (Simini et al., 2021), that may provide more accurate estimations using all available data at the cost of decreased transparency or increased model complexity.

5.1. Conclusion

To conclude, we have investigated the use of multiple types of data and models in estimating inter-regional mobility during the onset of the COVID-19 pandemic in Finland. The highest accuracy was provided by a model that combined past mobile phone data with public up-to-date road traffic data. The use of road traffic data also improved the performance of traditional static mobility models. These results pave the way for a data fusion approach to estimating inter-regional mobility flows.

Funding

A.P.L.C., T.A-N., L.L., and M.K. acknowledge funding from the project 105572 NordicMathCovid which is part of the Nordic Programme on Health and Welfare funded by NordForsk. J.S., T.H., and Z. H. acknowledge funding from the Strategic Research Council at the Academy of Finland (NetResilience consortium, Grant Nos. 345188 and 345183).

CRediT authorship contribution statement

Sara Heydari: Conceptualization, Methodology, Data curation, Investigation, Visualization, Software, Writing - original draft, Writing - review & editing. **Zhiren Huang:** Conceptualization, Methodology, Data curation, Investigation, Visualization, Software, Writing - original draft. **Takayuki Hiraoka:** Conceptualization, Methodology, Writing – review & editing. **Alejandro Ponce de León Chávez:** Data curation. **Tapio Ala-Nissila:** Conceptualization, Writing – review & editing, Project administration, Funding acquisition. **Lasse Leskelä:** Conceptualization, Methodology, Writing – review & editing, Project administration, Funding acquisition. **Mikko Kivelä:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Jari Saramäki:** Conceptualization, Methodology, Supervision, Funding acquisition, Project administration, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The calculations presented above were performed using computer resources within the Aalto University School of Science "Science-IT" project.

Appendix A. Python codes for implementing the models

To facilitate the reproducibility of this study, we are sharing the Python codes developed to implement our models (SaraHeydari, 2022).

Appendix B. Datasets

B.1. Road traffic

The road traffic data used in this study are provided by ITM Finland Ltd (Intelligent Traffic Management Finland). Vehicle count data are collected by over 450 induction loops installed into the road network. These sensors capture the number of vehicles passing by each induction loop in time intervals of five minutes, together with their average speed. The data are available from the year 1995 to the present day and can be downloaded from https://www.digitraffic.fi/.

We have taken part in implementing a Python package that allows to fetch the road traffic data for any desired time period and also to aggregate the information geographically or temporally. The package is accessible on Github (Fin-traffic data, 2021). In the present study, we have used the package to coarse-grain the traffic-count data to obtain the number of vehicles moving between neighbouring hospital districts during 6-h time bins. There are 20 hospital districts mainland Finland, as shown in Fig. S1).

B.2. Population statistics and commuting

The radiation and gravity models of Sections 3.4.1 and 3.4.2 are traditionally are used to estimate the number of commuters between different regions. These models estimate mobility based on the geographical distribution of the population. They often assume that number of commuters and jobs in an area are both proportional to the population on region. Here, instead, we use the actual distribution of commuters and jobs.

Data on the geographical distribution of jobs and workers in Finland are publicly available from Statistics Finland (Finland, 2021). The spatial resolution of the data is at the level of municipalities; there are in total 295 municipalities in mainland Finland, see Fig. S1). The data are available annually from 1987.

In this study, we use the statistics for the year 2017, which are the most recent data available at the time of this study. The publicly available data include the number of workers living in municipality i, which we denote by n_i, the number of jobs available in municipality j, which we denote by m_j, and the number of people working outside of their municipality of residence i, which we refer to as out-commuters and denote by n_{i+}.

These data are used as inputs data to the radiation and gravity models to estimate a static origin–destination matrix, to be augmented by the road traffic data as introduced in Section 3.4.3.

B.3. Mobile phone data

The mobile phone data set, licenced from the teleoperator Telia, is a set of time-stamped origin–destination matrices aggregated to the level of Finnish municipalities. The company has aggregated these matrices from base-station-level information on the numbers of users travelling between municipalities. A break of 20 min or less has been allowed during the trip. The numbers of trips have then been scaled by the company according to its market share ($\sim$ 30% Transport and Agency (2022)). Thus, assuming that the operator's customer base is representative of the population, the values should be a good proxy for the numbers of individuals travelling between regions. The specifications of the data are:

- Geographical resolution: Municipalities in Finland (total of 295 municipalities in mainland Finland)
- Time period: 2019.02.01–2019.05.31 and 2020.02.01–2020.05.31 (total of 8 months)
- Time resolution: 4 times a day (00–06, 06–12, 12–18, and 18–00)

The data are missing for some dates and origin municipalities due to low signal quality. We have removed those dates which are missing the trips related to more than 10 origins from the dataset (18 days in total) (See Table S1).

B.4. Publicly available mobility trend indices

In response to the COVID-19 pandemic, both Facebook and Google have published mobility trend indices that are derived from the usage of their mobile applications together with location services. These indices measure changes in mobility levels compared to the pre-pandemic baseline. It should be noted that it is hard to assess whether the users constitute a representative sample of the population, as no details on demographics are released.

B.4.1. Facebook movement range maps

The Facebook movement range maps project (Facebook, 2021) provides two kinds of mobility indices: *Change in Movement* and *Stay put*. In this study, we use the Change in Movement index which reports the overall level of mobility as compared to a pre-pandemic baseline period (i.e., Feb. 2020). This index is provided daily for the 5 regions of mainland Finland (see Fig. S1 for the geographical resolution).

B.4.2. Google community mobility reports

The Google movement trends (Maps, 2021) provide six mobility trend indices corresponding to different categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential. Similarly to the Facebook data, these indices report the changes in mobility compared to a pre-pandemic baseline. The indices are produced for 19 regions in mainland Finland once per day (see Fig. S1). In our models, we used the index related to transit stations as the proxy of changes in the intensity of inter-regional mobility.

Appendix C. An asymmetric radiation model for work commuting

Assume a spatial region (e.g. country) partitioned into a finite number of cells (e.g. municipalities) where cell i contains a set of resident travelers N_i (e.g. people with jobs) and a set of travel destinations M_i (e.g. workplaces). We denote size of sets N_i and M_i by n_i and m_i. For every traveler x and every travel destination y, there is a number W_{xy} describing how much x values destination y. The numbers W_{xy} are assumed distinct. We denote by $Z_{xj} = \max_{y\in M_j} W_{xy}$ the best value of cell j to traveler x. The destination cell of traveler x with residence in cell N_i is defined as

$$D_i(x) = \begin{cases} i, & \text{if } Z_{xi} = \max\limits_j Z_{xj}; \\ j, & \text{if } Z_{xj} > Z_{xi} > \max\limits_{k:d_{ik}<d_{ij}} Z_{xk}, \end{cases}$$

where d_{ij} is a cost (e.g. travel distance between suitably weighted geographic cell centers) associated with the ordered cell pair (i, j). Hence traveler x resident in cell i selects cell $j \neq i$ as destination if and only if j is the cell with least cost among the cells offering better value than Z_{xi}.

In the special case where $m_i = cn_i$ for some constant c, and d_{ij} is the Euclidean distance on $\mathbb{R}^2$, this model reduces to the radiation model described in Simini et al. (2012). In this paper, we too regard cost d_{ij} to be the Euclidean distance, but take the number of jobs and workers from official statistics rather than assuming a constant relation between m_i and n_i,

Proposition 5.1. Assume that the values W_{xy} are mutually independent and distributed according to a continuous probability distribution on $\mathbb{R}$ (atom-free property). Then the random variables $D_i(x)$ are mutually independent and distributed as

$$\mathbb{P}(D_i(x) = j) = \begin{cases} \dfrac{m_i}{\sum\limits_k m_k}, & j = i; \\ \dfrac{m_i m_j}{(m_i + s_{ij})(m_i + m_j + s_{ij})}, & j \neq i, \end{cases} \tag{20}$$

where $m_i = |M_i|$ and $s_{ij} = \sum_{k\neq i:d_{ik}<d_{ij}} m_k$.

Proof. (*Proof of Proposition 5.1*) (i) Fix a traveler x with residence cell N_i. The assumptions that the values W_{xy} are independent and distributed according to a continious probability distribution imply that there is almost surely a unique y^* for which $y \mapsto W_{xy}$ attains its maximum. By symmetry, each destination y attains the maximum equally likely. Hence the maximum is attained by a destination y in cell i with probability $\frac{|M_i|}{|\cup_k M_k|} = \frac{m_i}{\sum_k m_k}$. This is the probability that $D_i(x) = i$.

(ii) Denote $S_{ij} = \{k \neq i : d_{ik} < d_{ij}\}$, and let $Z_{x,ij} = \max_{k\in S_{ij}} Z_{x,k}$ and $M_{ij} = \cup_{k\in S_{ij}} M_k$. Note that for $j \neq i$,

$$\begin{aligned} \mathbb{P}(D_i(x) = j) &= \mathbb{P}(Z_{x,ij} \leqslant Z_{x,i} < Z_{x,j}) \\ &= \mathbb{P}(Z_{x,i} \geqslant Z_{x,ij}) - \mathbb{P}(Z_{x,i} \geqslant \max\{Z_{x,ij}, Z_{x,j}\}) \\ &= \mathbb{P}\left(\max_{y\in M_i} W_{xy} \geqslant \max_{y\in M_{ij}} W_{xy}\right) - \mathbb{P}\left(\max_{y\in M_i} W_{xy} \geqslant \max_{y\in M_j\cup M_{ij}} W_{xy}\right). \end{aligned}$$

The independence assumption and the atom-free property imply that there is an almost surely unique value $y \in M_i \cup M_{ij}$ for which $y \mapsto W_{xy}$ attains its largest value. Moreover, each of the random variables W_{xy} has an equal probability of being the largest. Therefore, the largest value belongs to M_i with probability $\frac{|M_i|}{|M_i\cup M_{ij}|} = \frac{m_i}{m_i+s_{ij}}$. Hence

$$\mathbb{P}\left(\max_{y\in M_i} W_{xy} \geqslant \max_{y\in M_{ij}} W_{xy}\right) = \frac{m_i}{m_i + s_{ij}}.$$

A similar reasoning shows that

$$\begin{aligned} \mathbb{P}\left(\max_{x\in M_i} W_x \geqslant \max_{x\in M_j\cup M_{ij}} W_x\right) &= \frac{|M_i|}{|M_i \cup M_j \cup M_{ij}|} \\ &= \frac{m_i}{m_i + m_j + s_{ij}}. \end{aligned}$$

Hence the claim follows. □ Based on Eq. 5.1, the probability that a worker living in region i also works in i is equal to $m_i/\sum_k m_k$. However, we can obtain the diagonal entries of OD-matrix -e.g. number of people who work in their regions of residence- from the official statistics in Finland (see Section B.2 for more details on the data). We combine the empirical information on the diagonal entries with the probabilities related to the non-diagonal entries based on Eq. 5.1. Then, the probability that a worker living in municipality i works in municipality j would be:

$$p_{ij} = \begin{cases} 1 - \dfrac{n_{i+}}{n_i}, & j = i; \\ c_i \dfrac{m_i m_j}{(m_i + s_{ij})(m_i + m_j + s_{ij})}, & j \neq i, \end{cases} \tag{21}$$

where $n_{i+} = n_i - n_{ii}$ is the number of out-commuters of i, $s_{ij} = \sum_{k\neq i:d_{ik}<d_{ij}} m_k$ and the normalization constant is given by $c_i = \frac{n_{i+}}{n_i}\left(1 - \frac{m_i}{n}\right)^{-1}$ with $n = \sum_i n_i$. This can be interpreted as follows: A person in municipality i flips a coin and with probability n_{i+}/n_i decides to be an out-commuter (work outside the home municipality). Conditionally on being an out-commuter, the target municipality of a worker living in municipality i is sampled from the

conditional law of $D_i(x)$ given $D_i(x) \neq i$, where $D_i(x)$ is distributed according to (20). If all n_i workers living in i choose their work destination using (21), then the mean number of people living in i and working in j equals

$\mu_{ij} = n_i p_{ij}$.

References

Balcan, D., Colizza, V., Gonçalves, B., Hud, H., Ramasco, J.J., Vespignani, A., 2009. Multiscale mobility networks and the spatial spreading of infectious diseases. Proc. Natl. Acad. Sci. U.S.A. 106 (51), 21484–21489.

Barbieri, D.M., Lou, B., Passavanti, M., Hui, C., Hoff, I., Lessa, D.A., Sikka, G., Chang, K., Gupta, A., Fang, K., Banerjee, A., Maharaj, B., Lam, L., Ghasemi, N., Naik, B., Wang, F., Mirhosseini, A.F., Naseri, S., Liu, Z., Qiao, Y., Tucker, A., Wijayaratna, K., Peprah, P., Adomako, S., Yu, L., Goswami, S., Chen, H., Shu, B., Hessami, A., Abbas, M., Agarwal, N., Rashidi, T.H., 2021. Impact of COVID-19 pandemic on mobility in ten countries and associated perceived risk for all transport modes. PLoS ONE 16 (2 February), 1–18. https://doi.org/10.1371/journal.pone.0245886.

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C.R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J.J., Simini, F., Tomasini, M., 2018. Human mobility: Models and applications. Phys. Rep. 734, 1–74.

Belik, V., Geisel, T., Brockmann, D., 2011. Natural Human Mobility Patterns and Spatial Spread of Infectious Diseases. Phys. Rev. X 1 (1), 1–5.

Brockmann, D., Theis, F., 2008. Money circulation, trackable items, and the emergence of universal human mobility patterns. IEEE Pervasive Comput. 7 (4), 28–35.

Do, L.N., Vu, H.L., Vo, B.Q., Liu, Z., Phung, D., 2019. An effective spatial-temporal attention based neural network for traffic flow prediction. Transp. Res. Part C 108, 12–28.

Facebook (2021), Movement range maps,https://dataforgood.facebook.com/dfg/tools/movement-range-maps. [Online; accessed 29-Sep.-2021].

Fin-traffic data (2021), Git repository for downloading road traffic data, https://doi.org/10.5281/zenodo.6543469. [Online; accessed 25-Nov.-2022].

Finland, S. (2021), Key figures on population by region, 1990-2020. http://pxnet2.stat.fi/PXWeb/pxweb/en/StatFin/StatFin_vrm_vaerak/statfin_vaerak_pxt_11ra.px/. [Online; accessed 12-Jan.-2021].

Finland, T.M. (2022), Data on number of vehicles on the roads in finland, https://www.digitraffic.fi/. [Online; accessed 12-Jan.-2022].

Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., Trasarti, R., 2011. Unveiling the complexity of human mobility by querying and mining massive trajectory data. VLDB J. 20 (5), 695–719.

Guirao, B., Campa, J.L., Casado-Sanz, N., 2018. Labour mobility between cities and metropolitan integration: The role of high speed rail commuting in spain. Cities 78, 140–154.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., Ratti, C., 2014. Geo-located Twitter as proxy for global mobility patterns. Cartogr. Geographic Inf. Sci. 41 (3), 260–271. https://doi.org/10.1080/15230406.2014.890072.

Huang, X., Li, Z., Jiang, Y., Li, X. and Porter, D. (2020), Twitter, human mobility, and COVID-19, arXiv.

Huang, Z., Ling, X., Wang, P., Zhang, F., Mao, Y., Lin, T., Wang, F.Y., 2018. Modeling real-time human mobility based on mobile phone and transportation data fusion. Transp. Res. Part C 96, 251–269. https://doi.org/10.1016/j.trc.2018.09.016.

Huang, Z., Wang, P., Zhang, F., Gao, J., Schich, M., 2018. A mobility network approach to identify and anticipate large crowd gatherings. Transp. Res. Part B 114, 147–170.

Kiashemshaki, M., Huang, Z. and Saramäki, J. (2022), Mobility signatures: A tool for characterizing cities using intercity mobility flows, Frontiers in Big Data 5. https://www.frontiersin.org/article/10.3389/fdata.2022.822889.

Kraemer, M.U., Yang, C.H., Gutierrez, B., Wu, C.H., Klein, B., Pigott, D.M., du Plessis, L., Faria, N.R., Li, R., Hanage, W.P., Brownstein, J.S., Layan, M., Vespignani, A., Tian, H., Dye, C., Pybus, O.G., Scarpino, S.V., 2020. The effect of human mobility and control measures on the COVID-19 epidemic in China. Science 368 (6490), 493–497.

Laharotte, P.A., Billot, R., Come, E., Oukhellou, L., Nantes, A., El Faouzi, N.E., 2015. Spatiotemporal analysis of bluetooth data: Application to a large urban network. IEEE Trans. Intell. Transp. Syst. 16 (3), 1439–1448.

Lenormand, M., Bassolas, A., Ramasco, J.J., 2016. Systematic comparison of trip distribution laws and models. J. Transp. Geogr. 51, 158–169.

Lenormand, M., Huet, S., Gargiulo, F., Deffuant, G., 2012. A universal model of commuting networks. PLOS ONE 7 (10), 1–7.

Lu, X., Bengtsson, L., Holme, P., 2012. Predictability of population displacement after the 2010 haiti earthquake. Proc. Natl. Acad. Sci. (USA) 109, 11576–11581.

Luca, M., Barlacchi, G., Lepri, B., Pappalardo, L., 2021. A survey on deep learning for human mobility. ACM Computing Surveys (CSUR) 55 (1), 1–44.

Maps, G. (2021), Community mobility report, https://www.google.com/covid19/mobility/. [Online; accessed 29-Sep.-2021].

Masucci, A.P., Serras, J., Johansson, A., Batty, M., 2013. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. Phys. Rev. E 88 (2), 1–8.

Provenzano, D., Hawelka, B., Baggio, R., 2018. The mobility network of European tourists: a longitudinal study and a comparison with geo-located Twitter data. Tourism Rev. 73 (1), 28–43.

Ravenstein, E., 1885. The Laws of Migration. J. Stat. Soc. London 48 (2), 167–235. URL: http://www.jstor.org/stable/297.

Ren, Y., Ercsey-Ravasz, M., Wang, P., González, M.C., Toroczkai, Z., 2014. Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. Nat. Commun. 5.

Riley, S., Eames, K., Isham, V., Mollison, D., Trapman, P., 2015. Five challenges for spatial epidemic models. Epidemics 10 (2015), 68–71.

SaraHeydari (2022), Saraheydari/dynamic-mobility: A repository containing the code for implementating the models described in this paper, https://doi.org/10.5281/zenodo.6586326. [Online; accessed 25-Nov.-2022].

Schlosser, F., Maier, B.F., Jack, O., Hinrichs, D., Zachariae, A., Brockmann, D., 2020. Covid-19 lockdown induces disease-mitigating structural changes in mobility networks. Proc. Nat. Acad. Sci. 117 (52), 32883–32890.

Simini, F., Barlacchi, G., Luca, M., Pappalardo, L., 2021. A deep gravity model for mobility flows generation. Nat. Commun. 12 (1), 1–13.

Simini, F., González, M.C., Maritan, A., Barabási, A.L., 2012. A universal model for mobility and migration patterns. Nature 484 (7392), 96–100.

Simini, F., González, M.C., Maritan, A., Barabási, A.-L., 2012. A universal model for mobility and migration patterns. Nature 484 (7392), 96–100. https://doi.org/10.1038/nature10856.

State, B., Weber, I., Zagheni, E., 2013. Studying inter-national mobility through IP geolocation. In: WSDM 2013 – Proceedings of the 6th ACM International Conference on Web Search and Data Mining, pp. 265–274.

Statistics in Finland, C. (2021), Number of confirmed coronavirus cases in finland, https://experience.arcgis.com/experience/92e9bb33fac744c9a084381fc35aa3c7. [Online; accessed 09-Dec.-2021].

Tian, H., Liu, Y., Li, Y., Wu, C.H., Chen, B., Kraemer, M.U., Li, B., Cai, J., Xu, B., Yang, Q., Wang, B., Yang, P., Cui, Y., Song, Y., Zheng, P., Wang, Q., Bjornstad, O.N., Yang, R., Grenfell, B.T., Pybus, O.G., Dye, C., 2020. The impact of transmission control measures during the first 50 days of the COVID-19 epidemic in China. Science 642 (May), 638–642.

Transport, F. and Agency, C. (2022), Mobile subscriptions, https://www.traficom.fi/en/statistics/mobile-subscriptions-0. [Online; accessed 24-Mar.-2022].

Vespignani, A., 2009. Predicting the behavior of techno-social systems. Science 325 (5939), 425–428.

Wang, P., Hunter, T., Bayen, A.M., Schechtner, K., González, M.C., 2012. Understanding road usage patterns in urban areas. Scientific Rep. 2 (1), 1–6.

Wilson, A.G., 1971. A family of spatial interaction models, and associated developments. Environ. Planning A 3 (1), 1–32.

Xie, P., Li, T., Liu, J., Du, S., Yang, X., Zhang, J., 2020. Urban flow prediction from spatiotemporal data using machine learning: A survey. Inf. Fusion 59, 1–12.

YLE, F.B.C. (2020a), Finland closes schools, declares state of emergency over coronavirus, https://yle.fi/news/3-11260062. [Online; accessed 07-Mar.-2022].

YLE, F.B.C. (2020b), Finland extends coronavirus emergency measures to mid-may, https://yle.fi/news/3-11283446. [Online; accessed 23-Mar.-2022].

YLE, F.B.C. (2020c), Finland shuts down uusimaa to fight coronavirus, https://yle.fi/news/3-11276242. [Online; accessed 07-Mar.-2022].

YLE, F.B.C. (2021a), 13.3 10:48 thl: Finland close to epidemic threshold, 155 cases confirmed, https://yle.fi/news/3-11307944. [Online; accessed 07-Mar.-2022].

YLE, F.B.C. (2021b), One year since finland's first confirmed covid-19 case, https://yle.fi/news/3-11762849. [Online; accessed 07-Mar.-2022].

Zhang, D., Huang, J., Li, Y., Zhang, F., Xu, C., He, T., 2014. Exploring human mobility with multi-source data at extremely large metropolitan scales. In: Proceedings of the 20th annual international conference on Mobile computing and networking, pp. 201–212.

Zhang, J., Zheng, Y. and Qi, D. (2017), Deep spatio-temporal residual networks for citywide crowd flows prediction, in Thirty-first AAAI conference on artificial intelligence.

Zipf, G.K., 1946. The P 1 P 2 D Hypothesis: On the Intercity Movement of Persons. Am. Sociol. Rev. 11 (6), 677.