Japanese Journal of Statistics and Data Science (2021) 4:763–781
https://doi.org/10.1007/s42081-021-00109-z

ORIGINAL PAPER

Statistics for High-Frequency Data

Anomaly detection of mobile positioning data with applications to COVID-19 situational awareness

Stefano Maria Iacus[1] · Francesco Sermi[1] · Spyridon Spyratos[1] · Dario Tarchi[1] · Michele Vespe[1]

Received: 13 October 2020 / Accepted: 23 January 2021 / Published online: 5 March 2021

Abstract

Due to an unprecedented agreement with the European Mobile Network Operators, the Joint Research Centre of the European Commission was in charge of collecting and analyze mobile positioning data to provide scientific evidence to policy makers to face the COVID-19 pandemic. This work introduces a live anomaly detection system for these high-frequency and high-dimensional data collected at European scale. To take into account the different granularity in time and space of the data, the system has been designed to be simple, yet robust to the data diversity, with the aim of detecting abrupt increase of mobility towards specific regions as well as sudden drops of movements. A web application designed for policy makers, makes possible to visualize the anomalies and perceive the effect of containment and lifting measures in terms of their impact on human mobility as well as spot potential new outbreaks related to large gatherings.

Keywords Anomaly detection · Data science · Mobile network operator data · High-dimensional time series analysis

1 Introduction

By means of a letter to European MNOs, the European Commission asked for fully anonymised aggregated mobility data. This represents an unprecedented case of business to government agreement during crisis time. In compliance with the 'Guidelines on the use of location data and contact tracing tools in the context of the COVID-19 outbreak' by the European Data Protection Board EDPB (2020), these data do not provide information about the behaviour of individuals; it can, however, give valuable insights into mobility patterns of population groups.

✉ Stefano Maria Iacus
stefano.iacus@ec.europa.eu

[1] European Commission, Joint Research Centre, Via Enrico Fermi 2749, 21027 Ispra, VA, Italy

The availability of Mobile Network Operators (MNO's) data at EU scale in almost real time has the potential to enhance the situational awareness about events deviations from "usual" mobility patterns. Such anomalies may identify large gatherings that could be used as input to meta-population modelling and early warning applications aiming at flagging and projecting clusters that may lead to increases of R_t, the reproduction number.

Despite the fact that mobility data alone cannot predict future needs, they can show already compelling citizens needs, like transportation or heathcare facility allocation needs and they represent well human behavior (Bwambale et al. 2020). Moreover, thanks to the capability of collecting mobile data at very high time frequency and space granularity, the time evolution of the mobility patterns can indeed show changes or ongoing trends or help to measure policy effects like the COVID-19 containment measures.

It is important to remark that, since mobile phone services unique subscribers[1] represent about 65% of the population across Europe (GSMA 2020), mobile data can reliably be used to capture the aggregate mobility patterns of the population.

In this work, we present an anomaly detection system for mobile positioning data data for 19, out of 27, member states of the European Union (namely: Austria, Belgium, Bulgaria, Czechia, Germany, Denmark, Estonia, Spain, Finland, France, Greece, Croatia, Hungary, Italy, Portugal, Romania, Slovakia, Sweden, Slovenia) plus Norway. The data have been provided for good, and within the scope of supporting the COVID-19 fight of the pandemic, by 17 different MNO's to the Joint Research Centre (JRC) of the European Commission. This work introduces a live anomaly detection system for these high frequency and high-dimensional data collected at European scale. Given the high volume and the diversity of the input data (see Sect. 2 for details), a robust system for anomaly detection was developed at JRC to detect not only excess of mobility but also sudden drops of mobility patterns.

As anomaly detection corresponds to structural change detection in time series of human mobility, they are one of the key elements of a COVID-19 monitoring and situational awareness system. As it is now recognized (Trafton 2020; Wong and Collins 2020), super-spreader events are often linked to gathering events (Szablewski et al. 2020; Honderich 2020; News 2020), which can be detected as deviations from expected mobility. Such anomalies can help retrospective analysis of emerging clusters or in ex-post contact tracing. At the same time, it is also well known that importation of cases is one of the main vehicles of virus spread in the early phases of the pandemic (Mouchtouri et al. 2020; Dickens et al. 2020; Pinotti et al. 2020; Chih 2020) especially when between two areas there is a high differential in the number of cases. This was the case of the Sardinia Island in Italy last summer, for which the system generated several inbound mobility anomalies in the first 10 days of August and originating from several other Italian regions with relatively high number of cases, leading to a surge of new cases at destination 10–14 days later. As the system detects also abrupt drops in mobility, became an extremely interesting tool for the

[1] All mobile services subscribers, including IoT, are about 86% of the population, 76% of which real smartphone users.

local authorities to monitor the extent and the response time to which the citizens adhere to the different containment measures like curfews, partial or full lock-downs in the second wave of the pandemic. Overall, the ability of detecting and quantifying the changes in human mobility can be used as input not only to epidemiological models and monitoring activities, but also to economic modelling to estimate the size of the causal effects of the different containment measures or the impact of non-pharmaceutical interventions on mobility. This work presents the general detection system and few insights leaving out all socio-economic and epidemiological analyses which will be part of further studies.

This work is structured as follows. Section 2 describes the input data in terms of volume, granularity and diversity. Section 3 describes the basic idea of the anomaly detection systems and its scopes and Sect. 4 shows examples of the output of the detection system. Section 5 presents some screenshots of an internal visualization platform aimed at policy makers. Section 6 summarizes the limits of the proposed approach.

2 The mobile positioning data

The data received by the JRC are in the form of ODM: Origin-Destination Matrix (Mamei et al. 2019; Fekih et al. 2020; Kishore et al. 2020). Although the concept is somehow known to the general public, it is important to describe their nature to justify why the anomaly detection system of Sect. 3 has to be designed simple yet robust to handle many different situations in a context of big and high frequency data.

To deliver their telecommunication services, the MNOs need to collect information details like, e.g., the customer's position, which needs to be constantly updated to route calls and data to the user. Two types of events are continuously being monitored: the Call Detail Records (CDR), which include mobile phone calls, messaging, and internet data accesses, and the eXtended Detail Records (XDR), which also include network signalling data.

As mentioned, the starting input data of the anomaly detection system of Sect. 3 is an ODM. Each cell $[i - j]$ of the ODM shows the overall number of *'movements'* (also referred to as 'trips' or 'visits') that have been recorded from the origin geographical reference area i to the destination geographical reference area j over the reference period.

To avoid any ex-post re-identification of individuals, before getting into an ODM, the data have to undergo several additional procedures such as deletion of any personal data, removal of singularities, thresholding, application of differential privacy (noise and distortions) methods and so forth. In fact, the ODM comes in fully anonymised and aggregated form so that the risk of re-identification of individuals is virtually impossible.

In general, an ODM (see also Fig. 1) contains the following minimal information: a timestamp for the *start time* and *end time* of the events considered, the areas of *origin* and *destination* and the *counts* (movements, trips, etc). The table below is a fictitious example of how data are received:

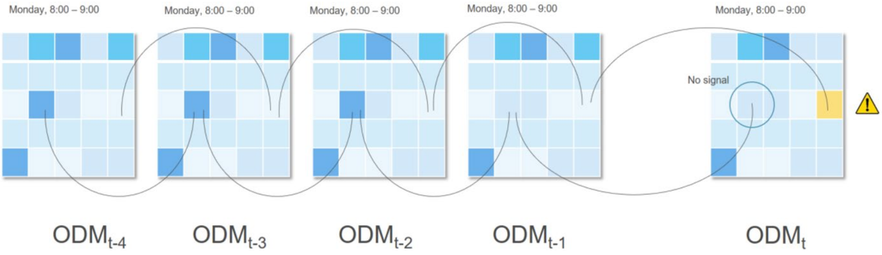

Fig. 1 The simplified logic behind the anomaly detection strategy: a sudden drop of the volume of the cell may identify an anomaly, while one within the natural variability of the data not

Origin_id	Destination_id	Start time	End time	Counts
zone001	zone001	2020-07-02 20:00:00	2020-07-02 21:00:00	3527
zone001	zone002	2020-07-02 20:00:00	2020-07-02 21:00:00	227
zone001	zone003	2020-07-02 20:00:00	2020-07-02 21:00:00	35
⋮	⋮	⋮	⋮	⋮

Sometimes, data may contain additional attributes like the distinction of type of contract (business, personal), age class, gender, national or foreign sim, etc, but in that case, the confidentiality threshold kicks in, i.e., most rows contain zeros to preserve privacy. As shown in the example above, the diagonal of the ODM contains the higher counts in our setup. To identify a movement between an origin area and a destination, it is necessary to define the *dwelling* (or stop) time. This dwelling time may vary from a few minutes to a few hours. A movement is recorded in the ODM only when the user stops for at least a duration equal to the dwelling time in the destination area having previously stopped for at least the same time in the origin area. An alternative way of defining a movement is to split the day in a number of time windows (normally 1-, 6- or 8-h long) and to count the users that move from one geographical area to another between time windows; in this case, a user's origin and destination areas are defined as those where the user spent most of the time in that time window. Also, consider that the number of movements do not correspond exactly to the number of persons, as under some circumstances, one person can generate more than one movement. Furthermore, the definition of geographical area can be very different from one case to another: it can be an administrative area or a regular spatial grid. The construction of the ODM therefore depends on a number of tuning parameters. Depending on the choice of these tuning parameters, an ODM will be able to capture some types of movements but not others. For instance, an ODM may capture movements that extend for a long period of time but not shorter movements and vice-versa. The technical details on how Call Detail Records (CDR) and eXtended Detail Records (XDR) are transformed into an ODM is specific to each MNO and it is considered proprietary technology not to be disclosed publicly although known to the authors.

Despite the diversity of the ODMs handled in this project, the ODM for a given MNO is consistent over time and relative changes are possible to be estimated.

Some applications of these data to different contexts than the one presented here can be found in Santamaria et al. (2020) and Iacus et al. (2020a, 2020b).

3 A simple, robust and scalable approach to anomaly detection

Detection of anomalies has a long history in statistics and quality control theory. In the context of change point analysis for the location parameter one can see, e.g., Bai (1997) and Csörgő and Horváth (1997) for i.i.d. settings and Bai (1994) for classical time series analysis, and in the context of the scale parameter for several classes of processes, e.g., Inclan and Tiao (1994) and Iacus and Yoshida (2012). These methods assume special data generation models and work with low-dimensional and low-frequency data mostly. In our case, we seek for robustness to data specification, computational efficiency and operational sustainability; therefore, several decisions have been made to simplify the approach.

On one side, the anomaly detection system has:

- to detect areas characterised by large increases of mobility that could be connected to gathering events;
- to systematically provide data-driven knowledge of such events that can be input to real epidemiological early warning systems.

on the other hand, the system has:

- to be computationally efficient given the dimensionality of the data in terms of frequency, spatial granularity and number of countries analysed;
- operationally feasible, i.e., produce almost real time and interpretable analysis;
- be robust with respect to high diversity of the input ODMs;
- be completely data driven in the sense that it should adapt itself to the time frame and granularity of the data.

In which terms the problem that the proposed system for anomalies detection has to consider consists in handling high-dimensional data? As said, the ODM are generated by different MNOs with different time frequency and space granularity: the ODM can be as large as[2] 9800×9800 entries time the 24 hourly sampling. The system should be able to capture anomalies of two types: the excess of volume and the sudden drop of volume as well as unexpected filling of some elements of the sparse ODM matrix at hand. It has to consider a non symmetric approach, as sudden drops and unexpected excesses are structurally different. Being counts, the zero is a natural lower limit for low volumes times series, while the upper limit should be determined through standard statistical ideas. We used a simple approach that takes into account both privacy thresholds (we do not consider cells whose moving average is below the

[2] For example, for Italy and many other member states.

threshold *th* (20 in our application[3]), natural variability and moving average. As it is well known that there exists both intra-daily, intra-weekly and seasonal patterns, we apply short period moving average from the given date, time frame and space granularity. Let *i* be the origin, *j* the destination, *s* the start time and *e* the end time of the sampling of the ODM for the date *d*. We denote each cell of the ODM matrix by

$$ODM^d_{s,e}(i,j)$$

where *i* and *j* spans the set of unique origin and destination labels $\mathcal{C}$, *d* is a calendar date and *s*, *e* are typically timestamps in the format "`YYYY-MM-DD HH:MM:SS`" though in our case they are in the order one or several hours. If we want to consider the total inbound flow to a cell *j*, we use the notation

$$ODM^d_{s,e}(\cdot,j) = \sum_{i\in\mathcal{C}} ODM^d_{s,e}(i,j)$$

and we denote by

$$ODM^d_{s,e}(i,\cdot) = \sum_{j\in\mathcal{C}} ODM^d_{s,e}(i,j)$$

the outbound movements from cell *i*. As there are situations in which the local movements are not interesting or such that the diagonal entries of the ODM matrix do not represent movements but people who stay in the same cell, we also consider the same quantities without the diagonals, i.e.,

$$\overline{ODM}^d_{s,e}(\cdot,j) = \sum_{i\in\mathcal{C},i\neq j} ODM^d_{s,e}(i,j)$$

and we denote by

$$\overline{ODM}^d_{s,e}(i,\cdot) = \sum_{j\in\mathcal{C},j\neq i} ODM^d_{s,e}(i,j).$$

The moving average is take over the previous *p* periods ($p = 4$ in our application), i.e.,

$$MA^d_{s,e}(i,j) = \frac{1}{p}\sum_{t=1}^{p} ODM^{d-t}_{s,e}(i,j)$$

and the rolling standard deviation is calculated similarly

$$SD^d_{s,e}(i,j) = \sqrt{\frac{1}{p}\sum_{t=1}^{p}\left(ODM^{d-t}_{s,e}(i,j)\right)^2 - \left(MA^d_{s,e}(i,j)\right)^2}$$

[3] The privacy threshold ranges from 5 to 20 across MNO, so we decided to keep a common value for all operators.

The reason why $p = 4$ has been chosen is due to the fact that human mobility have persistent patterns by day of the week, time of the day, season, etc. Taking $p = 4$ is equivalent to consider the past 4 weeks (i.e., the previous month) of data for the same time slot and day of week. Taking p smaller is statistically not reasonable as the moving average will be very unlikely to represent a trend; taking it larger will consider a too long time frame with the effect of smoothing out too much the seasonality. In any event, p is a tuning parameter that the researcher can fix according to his or her experience with the particular data ad hands.

In the event that for one or more past dates, the data are not available, the *MA* and *SD* are calculated on the available data only. If all past p data are missing, no signal will be estimated and the date d is marked as a "missing data" type. But historical variability in not enough as each ODM matrix, for different technical reasons at the MNOs level[4], may have a daily volume which is overall different from that of previous dates. This happens rarely but should be taken into account to avoid instrumental false signals, especially toward zero. Therefore, to take into account the overall variability, we select a first threshold Δ_q corresponding to the $q = 75\%$ quantile of the distribution of elements of the matrix $MA^d_{s,e}$ such that $MA^d_{s,e}(i,j) \geq th$. The *upper limit* is then set to

$$U^d_{s,e} = \max(MA^d_{s,e} + \Delta_q, MA^d_{s,e} + 3SD^d_{s,e}),$$

and the *lower limit* to

$$L^d_{s,e} = \min(MA^d_{s,e} - \Delta_q, MA^d_{s,e} - 3SD^d_{s,e}, 0).$$

The $MA^d_{s,e} + \Delta_q$ limit is larger than the limit $MA^d_{s,e} + 3SD^d_{s,e}$ (similarly for the lower limit) very rarely and this occurs only when there is a technical problems in the data. It is a sort of robust safeguard against extreme outliers that may occur for the technical reasons explained in the above.

In essence, the method has three tuning parameters: the confidentiality threshold th, the quantile level q of the distribution of the ODM and the number of past periods p. In our application, we have $(th, q, p) = (20, 0.75, 4)$.

Therefore, this is a simple 3-sigma approach combined with a robust evaluation of daily variability. More sophisticated time series approach or stochastic modelling (like inhomogeneous periodical Poisson process modelling) could have been used in spite of parametric tuning and estimation as well as computational time. Indeed, the present approach has been chosen also because of the need of the speed of calculation. All the formulas above have been implemented in R (R Core Team 2020) via sparse matrix linear algebra and, whenever possible, calculation on the data base have been used to reduce the data transfer bottleneck. The present approach can handle, for a single date, in less than an hour the analysis of 17 MNOs operators, providing data for 23 countries, at daily and, possibly, hourly frequencies. For example, for a single MNO operator for the country Italy, we have an ODM matrix of about 9800×9800 cells × and 25 time frames. The analysis is performed also on the

[4] It might happen that new antennas are installed in a given location, or an update of the mobile network occurs, and similar other very technical aspects.

9800 rows ($\overline{ODM}^{d}_{s,e}(\cdot,j)$) and 9800 columns separately ($\overline{ODM}^{d}_{s,e}(i,\cdot)$), considering the past 4 weeks as well (for the moving average calculation), i.e., the calculation of the anomalies is done on the non-null[5] $(9800 \times 9800 + 2 \times 9800) = 96,059,600$ times series taking into account the 25 time frames for 5 dates (the present and the past p dates).

3.1 Classification of the severity of signals

To help the policy makers in assessing the severity of the signals detected by the system, a simple classification scheme has been designed. The signals are then marked as "lower" and "upper" signals and their intensity is evaluated in terms of relative increment with respect to the moving average. As the moving average *per se* is not interesting to the policy maker, the signals are transformed into viable information through the relative increment. Let us denote this increment by

$$INC^{d}_{s,e}(i,j) = \left(\frac{ODM^{d}_{s,e}(i,j)}{MA^{d}_{s,e}} - 1 \right) \cdot 100\%$$

then, the level of the signal is classified as

- level 0 = no signal, i.e. $L^{d}_{s,e}(i,j) \leq DM^{d}_{s,e}(i,j) \leq U^{d}_{s,e}(i,j)$,
- level 1 if $INC^{d}_{s,e}(i,j) < 50\%$,
- level 2 if $50\% \leq INC^{d}_{s,e}(i,j) < 100\%$,
- level 3 if $INC^{d}_{s,e}(i,j) \geq 100\%$.

For both lower ($DM^{d}_{s,e}(i,j) < L^{d}_{s,e}(i,j)$) and upper ($DM^{d}_{s,e}(i,j) > U^{d}_{s,e}(i,j)$) signals as well as for the inbound and outbound timeseries $\overline{ODM}^{d}_{s,e}(\cdot,j)$ and $\overline{ODM}^{d}_{s,e}(i,\cdot)$. This type of filtering is helpful for the visual inspection of the thousands of signals appearing on a daily analysis.

A possible extensions of this approach could consider also the spatial information contained in the data as in this approach the entries of the cells are considered independently (the only way they area considered together is using the overall quantile of the matrix). This type of approach will be computationally quite hard to solve and requires additional *ad hoc* hypotheses according to the MNO source, country and granularity, which we prefer not to use at this stage. Moreover, the introduction of a correlation structure in the statistical model will compromise the use of extreme parallelization of the system.

Indeed, a future direction of research to take into account the spatial component is to include not the physical distance but the notion of connectivity between clusters of origins and destinations. Clustering will reduce the dimensionality and hence the computational burden, still loosing the ability to fully parallelize the processes, a compromise to be understood yet. Indeed, in a related study by the same authors, the notion of Mobility Functional Areas (MFAs) has been introduced (Iacus

[5] Although many of the cells of the ODM matrix are empty being a sparse matrix, in a single day several thousands of them are not null and, therefore, should be considered in the analysis.

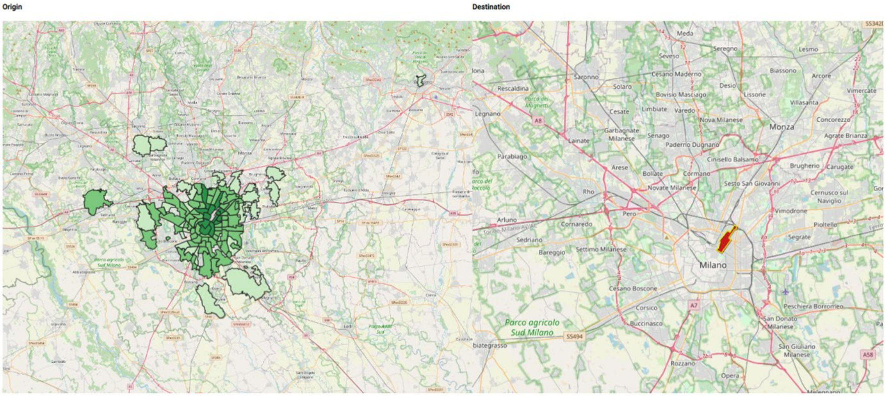

Fig. 2 Map of inbound signals in Milan, Central Station (ACE: 003015146C005). People arrive at the Central Station to participate to a public event. On the left, the (green) areas of origin for people terminating their trips to Central Station (red color). Movements are from around the city but also from external cities like Pavia and Bergamo. The intensity of the color is related to the relative amount of movements towards the destination

et al. 2020b). Each MFA represent a cluster of highly interconnected areas and it is obtained through network analysis by clustering the network graph generated by the daily ODMs. As these clusters vary daily, the subset of common subgroups of regions that appears at least 75% of the times in the same cluster are retained to form the final MFAs. The future work is to introduce this correlation structure into the anomaly detection system via Generalized Network AutoRegressive processes (GNAR) in the spirit of Knight et al. (2020) and Dahlhaus and Eichler (2003), but tailored for the high dimensionality of our setup.

The next section provides some examples of automatic detection in practice.

4 Examples of automatic detection: the case of Italy

As an example, we present graphically a few cases in which the system has spotted anomalies according to the previous description of the model. This section shows cases for which the system is expected to generate signals due to well-known public events, as well as cases that were unknown *ex-ante*. For privacy reasons, and because the scope of this system is not the tracking of people, we remove any information on the events and places that can potentially lead to identification of citizens or groups of people or communities, in accordance with the GSMA privacy guidelines[6] for COVID-19 data sharing and the Letter of Intent between the European Commission and the GSMA.

[6] https://www.gsma.com/publicpolicy/wp-content/uploads/2020/04/The-GSMA-COVID-19-Privacy-Guidelines.pdf

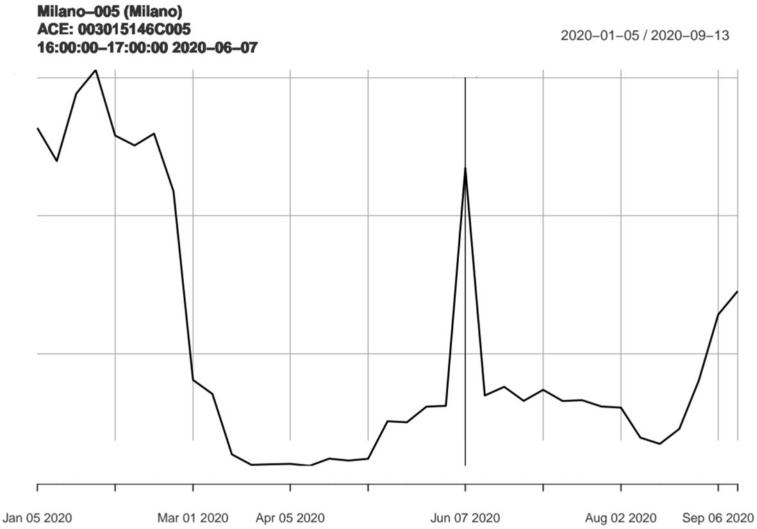

Fig. 3 Inbound signal in Milan, Central Station (ACE: 003015146C005), 4–5 p.m., on 7th June 2020. Excess of +250% with respect to the previous 4 weeks. People going home from the public event?

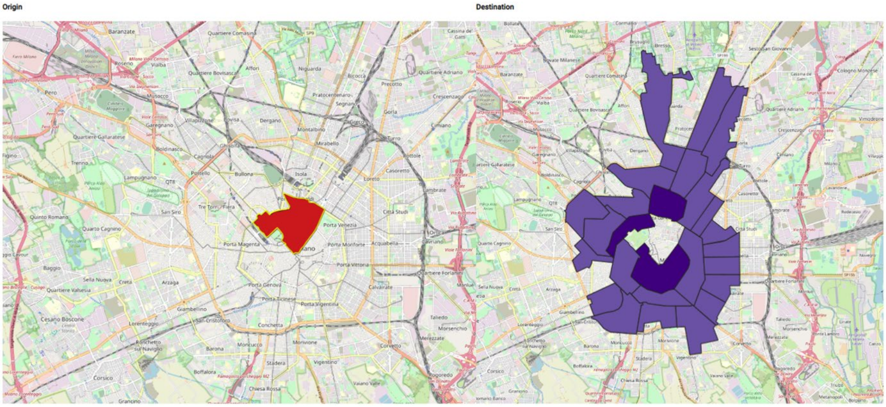

Fig. 4 Outbound hourly signals from the area of the gathering near the US General Consulate in Milan on 7th June 2020 between 6 and 7 p.m. People going home after the event?

4.1 Public event in Milan, 2020-06-07

This is an example of expected gathering due to a public event in the city of Milan, Italy, and it serves as a benchmark to assess the correctness of the anomaly signals. The system spotted anomalies in daily and hourly inbound flows on 7th June 2020,

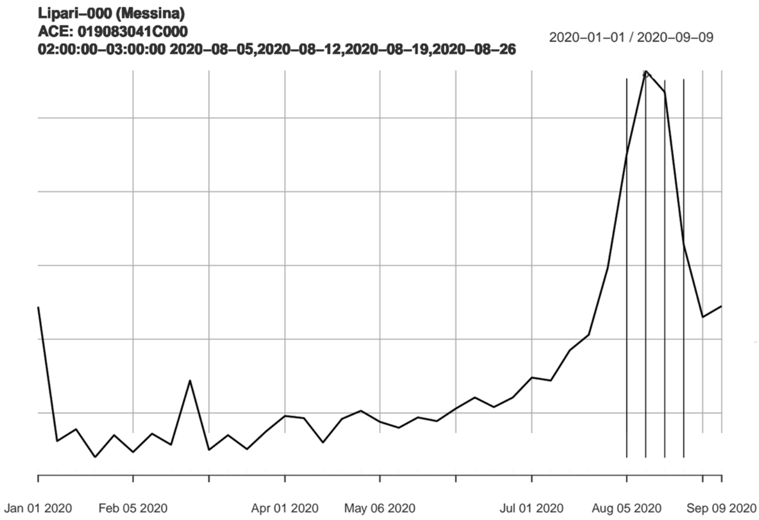

Fig. 5 Inbound signal in Lipari at night (ACE: 003015146C005), 2 a.m.–3 a.m., possibly connected to Summer nightlife and tourism activities. Excess of about + 300%

compared to the previous 4 weeks, to the census cell ACE[7]: 003015146C005, which corresponds to the Central Railways Station in the city of Milan. Figure 2 shows the daily inbound movements for this ACE on a map while Fig. 3 shows a peak of inbound hourly traffic between 4 and 5 p.m.

Similarly, an outbound signal for the area where the gathering actually took place is shown in Fig. 4. Clearly we can only guess that this signals are generated by specific gatherings, but the time series plot of Fig. 3 clearly shows an unexpected peak compared to the history of the mobility from and to that geographical area. A likely story for these signals is that people gather during the morning to Milan also from outside, then after the events either go home with train (around 5–6 p.m.) or by local transportation means to closer places (around 6–7 p.m.).

4.2 Summer nightlife and tourism: Lipari (Sicily), mid to end of August 2020

A different anomaly increase in mobility pattern has been spotted for inbound flows to the census cell ACE: 019083041C000, which corresponds to island of Lipari (Sicily), around mid to end of August 2020 at 2–3 a.m. as shown in Fig. 5. These events have been reported to increase the number of infected people, similar to the case of Sardinia on the same dates.

4.3 Venice Carnival (8–25 Feb 2020)

The system spotted anomalies increases also in inbound flows to the two census cells (ACE: 005027042C001 / 005027042C002), which corresponds to the city of

[7] ACE is the smaller census administrative geographical unit for Italy.

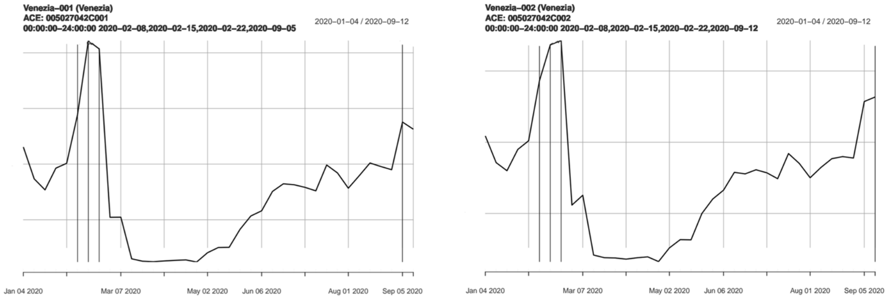

Fig. 6 Inbound daily signals in the two ACE cells of Venice, during the events of Carnival 2020, 8, 15 and 22 February 2020

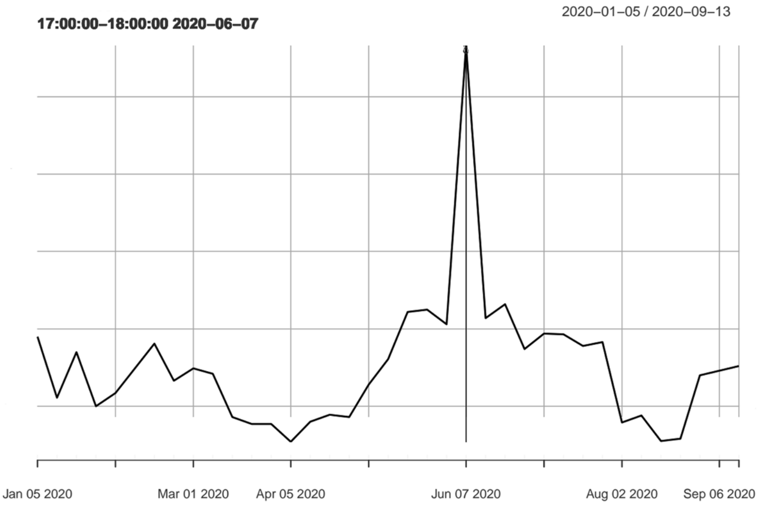

Fig. 7 Peak of movements toward a specific area around 5–6 p.m. Excess of about +300%

Venice. The dates are around the Carnival 2020 (18–22 February 2020) as shown in Fig. 6. These events are quite relevant in terms of predicting the pandemic as at that time no physical distancing measures were in place. The dates spotted by system are: 8, 15 and 22 February 2020.

4.4 An unexpected large gathering on 7th June 2020

While the previous anomalies could have been guessed, the system also usually finds several other cases. Just as an example, on June 7th, the system spotted and anomalous number of movements toward a very specific destination area (see

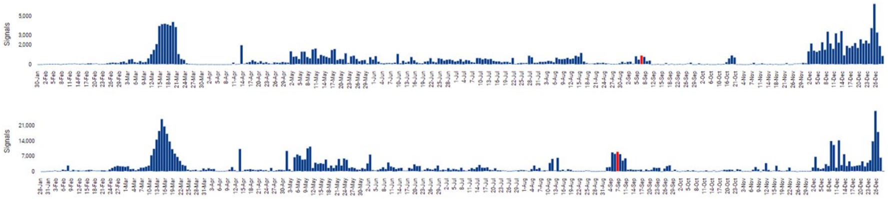

Fig. 8 The overall number of signals around the European countries. In the top panel the total number of daily signals and below the total number of hourly signals for the same time period. The time span of the two plots is not the same as for some MNOs we do not have hourly data and viceversa. The red bar represents the data for which the dashboard will show the signals on the map. In the example, it is set to 7 September 2020

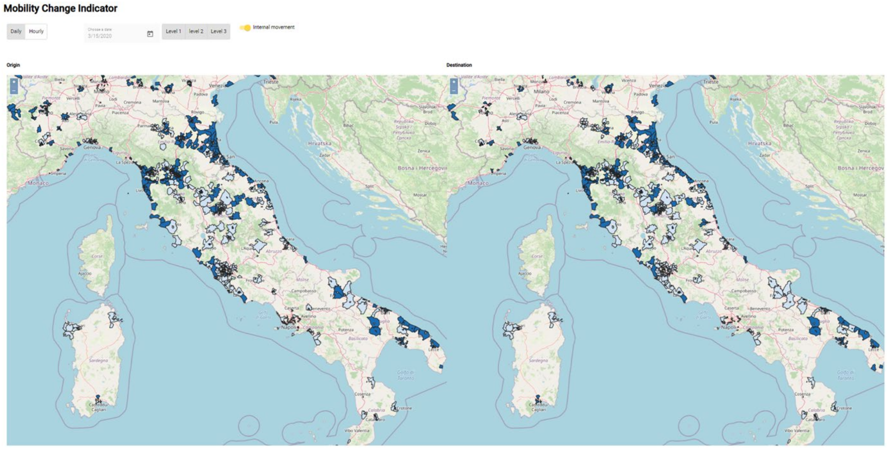

Fig. 9 Inbound and outbound signals on 16-03-2020, daily data. Blue = lower limit signal, Red = upper limit signal. Mobility almost stopping compared to the 4 previous weeks

Fig. 7). An increase of about 300% more movements than usual where registered between 5 and 6 p.m. This was quite surprising as it occurred few days after a national lockdown was lifted. If an increase in the number of cases were spotted around that location, and luckily this was not the case, the local authorities could have informed the local community to undergo COVID-19 testing implementing contact tracing that otherwise could have been impossible to do.

5 Visual application to explore anomalies aimed at policy makers

The previous set of examples show that expected and unexpected signals can be captured well by the system. The performance of the systems is essentially the same in all countries and for all mobile network operators considered. But while the

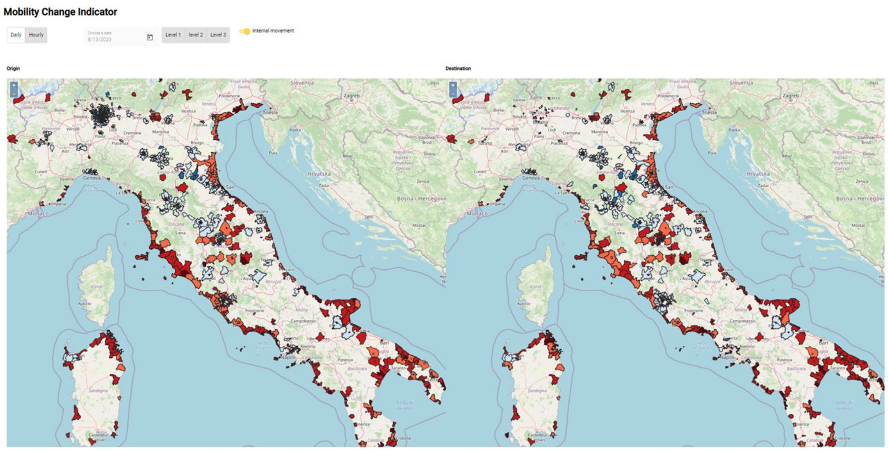

Fig. 10 Inbound and outbound signals on 13-08-2020, hourly data. Blue = lower limit signal, Red = upper limit signal. Mobility almost exploding in the coastal areas and still freezing ìn the cities, compared to the 4 previous weeks

time-series plot and the absolute numbers are of interest mainly for the modellers in epidemiology, the policy makers need a more intuitive tool to extract the same information. For this reason, an interactive web application to explore the signals has been developed and its in use by some policy makers.

A couple of examples for daily and hourly signals is given in Figs. 9 and 10. In Figs. 9 and 11, it is shown how mobility suddenly reduces after the lockdown which happened almost at the same time overall Europe around 16th March 2020. On the contrary, while Fig. 10 shows movements mainly towards, and in, the coastal areas around Italy and a reduction of mobility within the cities. This is expected and suggests that probably a seasonal effect should be take in consideration within the model, but this requires at least 2 complete years of data. Unfortunately, data from MNO are available from at most January 2020.

Figure 12 shows that the mobility in France and Spain was already restarted by mid May contrary to Italy. This clearly shows how the system can detect anomalies irrespective of the country and MNO data but it also enable the policy maker to have a broader view of the overall situation on a EU scale and the impact of containment and lifting measures.

Finally, Fig. 13 shows the overlap of the anomaly detection maps with the official ECDC data about number of cases per 100K inhabitants which serves to both the policy makers and the modelers to draw conclusions about the pandemic evolution and its relationship with human mobility.

Figure 8 shows the interactive daily and hourly histograms that are available in the visual dashboard. The reader can have an intuition of the amount of signals that

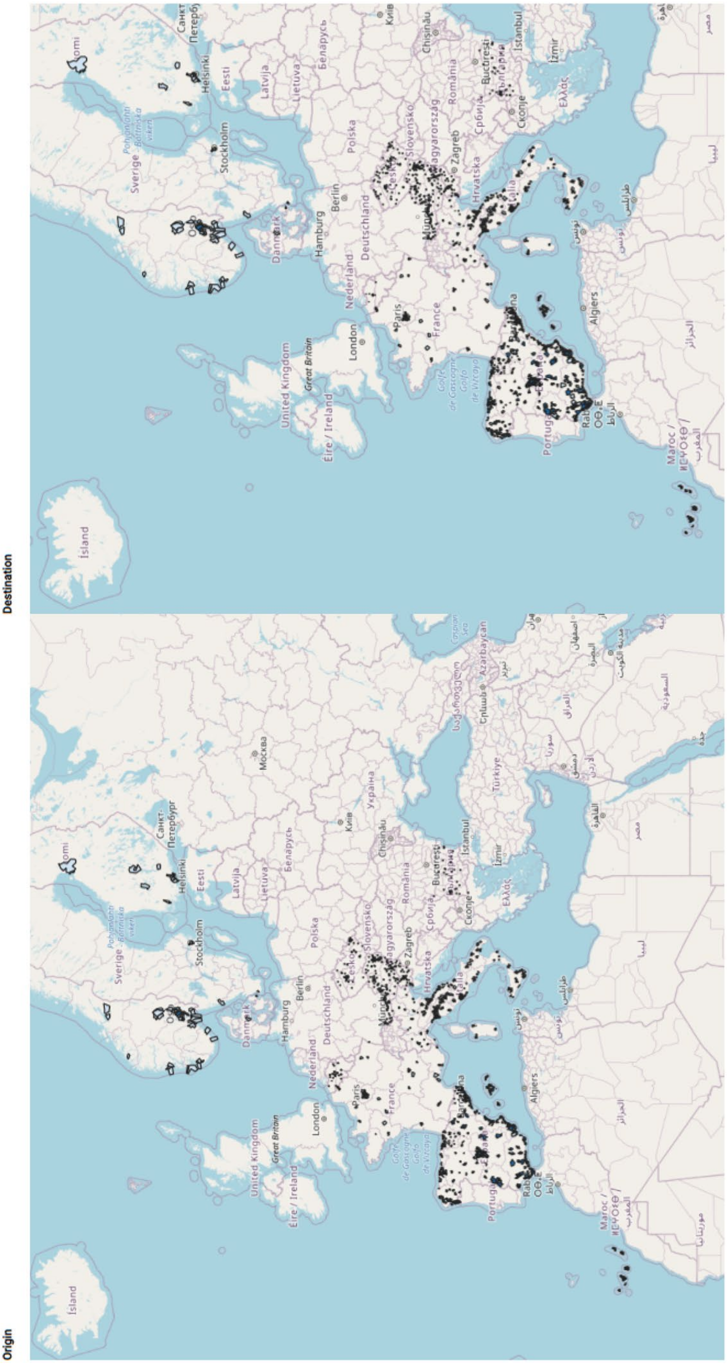

Fig. 11 Inbound and outbound signals on 16-03-2020, daily data for the whole set of European countries. Blue = lower limit signal, Red = upper limit signal. Mobility almost stopping compared to the 4 previous weeks in all countries

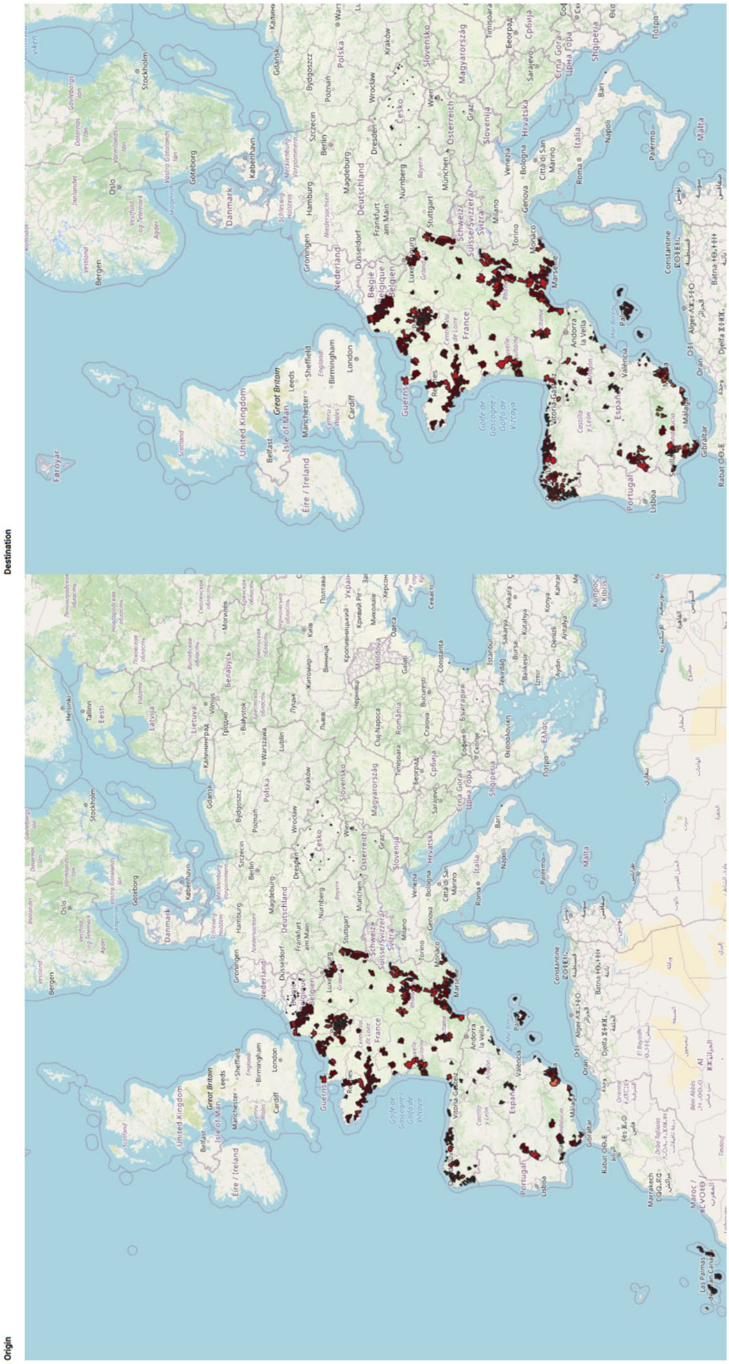

Fig. 12 Inbound and outbound signals on 16-05-2020, hourly data. Blue = lower limit signal, red = upper limit signal. Mobility re-exploding in Spain and France but not, e.g., in Italy compared to the 4 previous weeks. For Italy, we need to wait till July/August 2020. See also Fig. 10

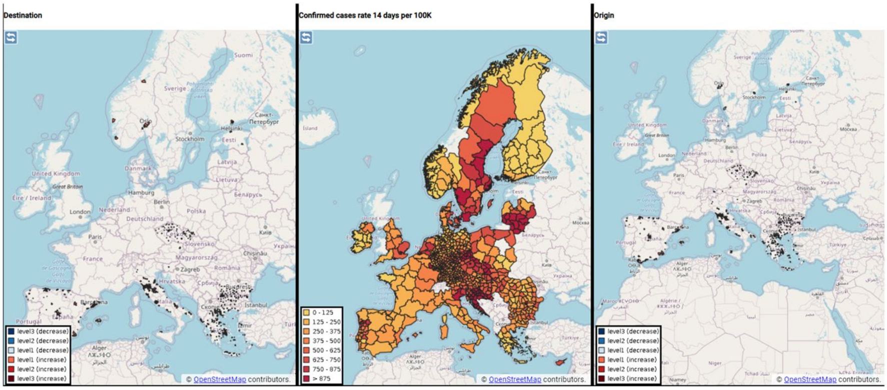

Fig. 13 The anomaly detection dashboard with the additional layer of official ECDC data on the number of cases per 100K inhabitants

are detected in almost real time[8] for this very large scale problem. The system has detected peaks up to 21,000 hourly signals within a day, or 5000 daily signals and sees from the histograms. In practice, the policy makers look at their own regions and filter the signals according the severity to monitor the effects of the implementation of the policy measures.

6 Conclusions and limits of this approach

As said, this simple and direct approach to the anomaly detection does not consider the spatial information contained in the data. This can be a nice addition in future developments of the system. Indeed, parametric and non-parametric geo-statistical models can also be considered at the cost of putting assumptions on the data (by country and MNO) and demanding for more computational time. The dimensionality of the problem is so high that, even using some restrictions like local dependency structure, it will become quite unfeasible to obtain model estimates in practical times though as parallel computing for millions of time series trajectories of each origin-destination dyad in the ODMs will be no longer an option. As mentioned in Sect. 3, future direction of investigation will consider the introduction into the anomaly detection system of Generalized Network AutoRegressive processes (GNAR) (2003; Knight et al. 2020).

The system has been designed to alert on mobility anomalies for early warning capacity in case of COVID-19 outbreaks. Since these anomalies can be generally attributed to large gatherings and unusual mobility patterns in a broader sense, the system is a precious tool to understand the potential spread of the virus in case of outbreaks. At the same time, the system can allow authorities to monitor the implementation of mobility restrictions.

[8] The system runs two times a day and whenever new MNO data are ingested into the infrastructure.

The system is not designed to be a tracking system as it is totally agnostic to reality. The examples of the previous sections show the validity of the system in case of a known event, and its detection serves to benchmark testing.

It is also worth to mention that the system has not be designed to produce a real COVID-19 early warning system but only to spot anomalies in the data in the terms explained in Sect. 3. This means that there is no direct link in this application between, e.g., the large gatherings spotted and the reproduction rate R_t of the COVID-19 pandemic. Our data could only serve as an input to further epidemiological models or to policy makers to assess the effectiveness of the containment measures.

Despite its limitations, the system seems to be able to capture what it is supposed to capture. It is fast to execute and can accommodate different sources of MNO data without any stringent assumptions rather than the confidentiality threshold $th = 20$, the length of the moving average $p = 4$ and the quantile level 75%. These are the only three tuning parameters of the anomaly detection system and can be changed by the researcher.

Acknowledgements The authors acknowledge the support of European MNOs (among which 3 Group - part of CK Hutchison, A1 Telekom Austria Group, Altice Portugal, Deutsche Telekom, Orange, Proximus, TIM Telecom Italia, Telefonica, Telenor, Tele2, Telia Company and Vodafone) in providing access to aggregate and anonymised data, an invaluable contribution to the initiative. The authors would also like to acknowledge the GSMA$^{(}$9) , colleagues from Eurostat $^{(}$10) and ECDC $^{(}$11) for their input in drafting the data request. Finally, the authors would also like to acknowledge the support from JRC colleagues, and in particular the E3 Unit, for setting up a secure environment to host and process of the data provided by MNOs, as well as the E6 Unit (the "Dynamic Data Hub" team) for their valuable support in setting up the data lake. The authors also acknowledge the Editor of the Special Issue and two anonymous referees for their constructive comments that allowed us to improve vastly the presentation of this study

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interests.

References

Bai, J. (1994). Least squares estimation of a shift in linear processes. *Journal of Times Series Analysis,15*, 453–472.

Bai, J. (1997). Estimation of a change point in multiple regression models. *The Review of Economics and Statistics,79*, 551–563.

Bwambale, A., Choudhury, C., Hess, S., & Iqbal, M. S. (2020). Getting the best of both worlds: a framework for combining disaggregate travel survey data and aggregate mobile phone data for trip generation modelling. *Transportation*. https://doi.org/10.1007/s11116-020-10129-5.

Chih-Yen, L., Wen-Hung, W., Aspiro, N., Sung-Pin, U., Po-Liang, T., Yen-Hsu, L., Ming-Lung, C., Seng-Fan, WY. (2020). Importation of SARS-CoV-2 infection leads to major COVID-19 epidemic in Taiwan. *International Journal of Infectious Diseases 97*, 240–244.

Csörgő, M., & Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. New York: Wiley.

Dahlhaus, R., Eichler, M. (2003). Causality and graphical models in time series. In Richardson (Eds.) *Highly structured stochastic systems*. University Press.

Dickens, B. L., Koo, J. R., Lim, J. T., Sun, H., Clapham, H. E., Wilder-Smith, A., Cook, A. R.: (2020). Strategies at points of entry to reduce importation risk of COVID-19 cases and reopen travel. *Journal of Travel Medicine* taaa141. 27(8). https://doi.org/10.1093/jtm/taaa141

EDPB. (04/2020). Guidelines 04/2020 on the use of location data and contact tracing tools in the context of the COVID-19 outbreak. https://edpb.europa.eu/our-work-tools/our-documents/linee-guida/guidelines-042020-use-location-data-and-contact-tracing_en. Accessed 1 Dec 2020.

Fekih, M., Bellemans, T., Smoreda, Z., Bonnel, P., Furno, A., & Galland, S. (2020). A data-driven approach for origin-destination matrix construction from cellular network signalling data: A case study of Lyon region (France). *Transportation*. https://doi.org/10.1007/s11116-020-10108-w.

GSMA. (2020). The mobile economy 2020 report. https://www.gsma.com/mobileeconomy/. Accessed 1 Dec 2020

Honderich, H. (2020). Coronavirus: What makes a gathering a 'superspreader' event? https://www.bbc.com/news/world-us-canada-53273382. Accessed 1 Dec 2020

Iacus, S. M., Santamaria, C., Sermi, F., Spyratos, S., Tarchi, D., Vespe, M. (2020a) Human mobility and COVID-19 initial dynamics. *Nonlinear Dynamics*. *101*, 1901–1919 (2020).

Iacus, S. M., Santamaria, C., Sermi, F., Spyratos, S., Tarchi, D., Vespe, M. (2020b). Mapping mobility functional areas (MFA) using mobile positioning data to inform COVID-19 policies JRC121299. https://ec.europa.eu/jrc/en/publication/mapping-mobility-functional-areas-mfa-using-mobile-positioning-data-inform-covid-19-policies, https://doi.org/10.2760/076318.

Iacus, S. M., & Yoshida, N. (2012). Estimation for the change point of the volatility in a stochastic differential equation. *Stochastic Processes and Their Applications,122*, 1068–1092.

Inclan, C., & Tiao, G. (1994). Use of cumulative sums of squares for retrospective detection of change of variance. *Journal of the American Statistical Association,89*, 913–923.

Kishore, N., Kiang, M., Engø-Monsen, K., Vembar, N., Balsari, S., Buckee, C. (2020). Mobile phone data analysis guidelines: Applications to monitoring physical distancing and modeling COVID-19. *OSF Preprints*. https://doi.org/10.1016/S2589-7500(20)30193-X.

Knight, M., Leeming, K., Nason, G., & Nunes, M. (2020). Generalized network autoregressive processes and the gnar package. *Journal of Statistical Software, Articles,96*(5), 1–36.

Mamei, M., Bicocchi, N., Lippi, M., Mariani, S., & Zambonelli, F. (2019). Evaluating origin-destination matrices obtained from CDR data. *Sensors,19*, 1440.

Mouchtouri, V. A., Bogogiannidou, Z., Dirksen-Fischer, M., Tsiodras, S., & Hadjichristodoulou, C. (2020). Detection of imported COVID-19 cases worldwide: early assessment of airport entry screening, 24 January until 17 February 2020. *Tropical Medicine and Health,48*(1), 79.

News, B. (2020). White house hosted COVID 'superspreader' event, says dr. fauci. https://www.bbc.com/news/election-us-2020-54487154. Accessed 1 Dec 2020.

Pinotti, F., Di Domenico, L., Ortega, E., Mancastroppa, M., Pullano, G., Valdano, E., Boëlle, P.-Y., Poletto, C., Colizza, V. (2020). Lessons learnt from 288 COVID-19 international cases: importations over time, effect of interventions, underdetection of imported cases. *medRxiv*. https://doi.org/10.1101/2020.02.24.20027326.

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Santamaria, C., Sermi, F., Spyratos, S., Iacus, S. M., Annunziato, A., Tarchi, D., et al. (2020). Measuring the impact of COVID-19 confinement measures on human mobility using mobile positioning data. A European regional analysis. *Safety Science,132*, 104925.

Szablewski, C., Chang, K., Brown, M., Chu, V., Yousaf, A., Anyalechi, N., et al. (2020). Sars-cov-2 transmission and infection among attendees of an overnight camp–Georgia, June 2020. *Morbidity and Mortality Weekly Report,69*, 1023–1025. https://doi.org/10.15585/mmwr.mm6931e1.

Trafton, A. (2020). Covid-19 "super-spreading" events play outsized role in overall disease transmission. https://news.mit.edu/2020/super-spreading-covid-transmission-1102. Accessed 1 Dec 2020.

Wong, F., & Collins, J. J. (2020). Evidence that coronavirus superspreading is fat-tailed. *Proceedings of the National Academy of Sciences, 117*(47), 29416–29418.