communications physics

Article

https://doi.org/10.1038/s42005-024-01909-x

Distorted insights from human mobility data

Riccardo Gallotti[1] ✉, Davide Maniscalco[1,2], Marc Barthelemy[3,4] & Manlio De Domenico[1,5,6] ✉

The description of human mobility is at the core of many fundamental applications ranging from urbanism and transportation to epidemics containment. Data about human movements, once scarce, is now widely available thanks to new sources such as phone call detail records, GPS devices, or Smartphone apps. Nevertheless, it is still common to rely on a single dataset by implicitly assuming that the statistical properties observed are robust regardless of data gathering and processing techniques. Here, we test this assumption on a broad scale by comparing human mobility datasets obtained from 7 different data-sources, tracing 500+ millions individuals in 145 countries. We report wide quantifiable differences in the resulting mobility networks and in the displacement distribution. These variations impact processes taking place on these networks like epidemic spreading. Our results point to the need for disclosing the data processing and, overall, to follow good practices to ensure robust and reproducible results.

In the past two decades, access to the rich high-resolution datasets was exclusive to a rather limited research community, which nevertheless produced a series of fundamental results about the statistical features of mobility behavior[1–7]. The potential impact of human mobility studies is enormous, as exemplified by most COVID-19 studies which rely heavily on mobility data[8,9]. The recent need for tackling promptly the problems associated with this pandemics both from the epidemiological and socio-economical perspective has been answered by many IT companies, which activated a number of 'Data For Good' programs and collaborations. This unprecedented access to mobility data by a large number of research groups led to the publication of a series of fundamental results about human behavior and, more recently, dynamics of interest such as the spreading of COVID-19[10–17] (see Supplementary Table I). However, even if this sudden abundance of datasets provided scientists with numerous new research opportunities, it also comes with new challenges. Without any doubt, an easier access to data favored the possibility of performing innovative data-informed analysis, in some cases across multiple countries and with a high temporal resolution. These recent studies also were in many cases fully reproducible, providing to the community both their codes and datasets[12,18–20]. However, for companies to be able to provide large amounts of data in almost real time, they had to overcome multiple limitations, and since the methods used are proprietary, the data pipeline cannot be openly released. As a consequence, key details might not be disclosed, often leaving to the final user the task to use data under uncertain, or even unknown, gathering methodology. Consequently, the lack of full knowledge about the measurement and processing details might prevent scientists from correctly interpreting the outcome of their analysis. Moreover, not being able to be in control of the data elaboration from the raw to their final, analyzable, form, limits the ability of researchers to tailor the data to the research question they have in mind. Even more crucially, each data provider extracts information using a different gathering and processing methodology, raising the question to which extent different datasets are in agreement with each other when used to model human mobility and derived dynamical processes, such as traffic or epidemics. The implicit assumption made is that the characteristics of the shared data is marginally affected by the details of the underlying methodology. Such an assumption is often explicitly acknowledged as a study limitation[11,20,21], but it has been never systematically tested or verified, and could in principle jeopardize the relevance of the obtained results.

Here, we illustrate both the opportunities provided and the limitations faced while studying mobility data coming from seven different providers (including both openly accessible and closed data): Google, Facebook, GPS-based applications (Cuebiq, Safegraph), Mobile phones, vehicle GPS blackboxes, and public census. We first provide a short description of these datasets and more details and discussions—in particular their limitations—can be found in the "Methods" section.

[1]Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo (TN), Italy. [2]Sorbonne Université, INSERM, Institut Pierre Louis d' Épidémiologie et de Santé Publique, IPLESP, Paris, France. [3]Université Paris-Saclay, CNRS, CEA, Institut de Physique Théorique, 91191 Gif-sur-Yvette, France. [4]CAMS (CNRS/EHESS), 54 Boulevard Raspail, 75006 Paris, France. [5]CoMuNe Lab, Department of Physics and Astronomy "Galileo Galilei", University of Padua, Padova, Italy. [6]Istituto Nazionale di Fisica Nucleare, Sez. Padova, Padova, Italy. ✉e-mail: rgallotti@fbk.eu; manlio.dedomenico@unipd.it

Results

Overview of the data

We considered a diverse set of mobility datasets, accounting for the movements of over 500 million individuals in 145 countries (see Supplementary Table II). Specifically, we used census data which provide statistical information about the inter-urban commuters' flows. We collected and analyzed official census data for the US at the county level[22] and for Italy aggregated at the province level[23]. We also used data coming from mobile phone data call detail records (CDR). We reconstruct the displacement distribution function for the US using published fitting parameters[1], and those for Portugal, Spain and France from a dataset published for reproducibility purposes[24]. Another source of mobility data comes from small blackboxes equipped with GPS trackers and accelerometers that are installed in private vehicles for insurance reasons and record trajectory data, and here we analyzed the urban and inter-urban displacement statistics of Italian drivers as in ref. 4. We studied the Google datasets that are elaborated from the opt-in service 'Google location history', where the trajectory data captures the movement of Android smartphone users between pairs of stop locations. We used the displacement statistics released by the authors of ref. 20, that covers almost all countries worldwide. Another fundamental source of data is provided by Facebook Data for Good Program[25] and obtained by tracking the movements of mobile phone users that opted-in to the Location History and Background Location collection services[26]. Here, we use the mobility network where the nodes represent administrative areas and the edges correspond to the long-range mobility flows between these areas. The flows are aggregated every 8 h and we further aggregated them on a weekly basis. If a flow is below a certain threshold (that we estimate being of order 10 users), it is not reported in the data. For epidemic simulations, we also used, as a proxy for population, the baseline values describing the number of Facebook active users in a given node, similarly averaged over a whole week (see "Methods"). The fifth dataset studied here comes from the Data for Good program of Cuebiq Inc.[27], which collects mobility data of anonymized mobile app users who opted-in to a large number of different location-based services in different countries including the US, the UK, Italy, Spain, France and Germany. Here, we also use the Cuebiq-HDR mobility data for Italy elaborated by Pepe and collaborators from the device-level data, and aggregated weekly at the province scale[28]. We also use 'mobility insight' flows describing the movements across the US, computed directly by Cuebiq and provided us within the framework of the Data4Good program. The flows are here also aggregated weekly and at the county scale. We computed the baseline flows by aggregating them over the weeks preceding the beginning of lockdowns in Italy and the US, rescaled the flows by the country total population and used the associated census population for epidemic modeling. Finally, the last dataset used in our study is provided by Safegraph and elaborated by Kang and collaborators[29]. This dataset collects statistics about visiting patterns of different points of interests (PoI) by aggregating anonymized location data from mobile applications and covers the US only, where the home location is identified for each user at the level of census block group. Here, we will use the flows aggregated weekly and at the county level[29], and a 2-month period before the US lockdown to define the baseline flows.

The dataset dependence of the displacement distribution

We first consider the distribution $P(L)$ of displacement distances L computed with various datasets. This distribution has been intensively studied during the last 15 years[1,2,4,20,30]. Studying $P(L)$ allows for comparing our results with published data where the analysis has been done on the raw data[4,31,32]. The tail behavior of $P(L)$ varies across countries[20], as it is product of the combination of mobility acting at multiple scales[4,5] while also combining individual contributions that differ by both gender[5] and socio-economic

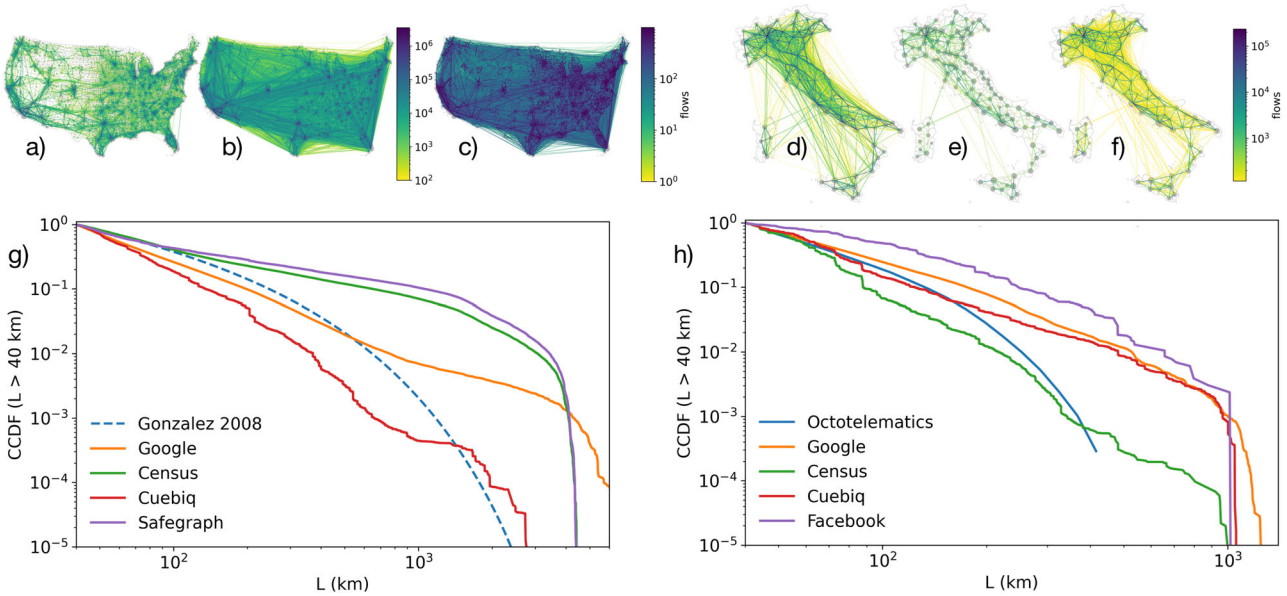

Fig. 1 | The baseline mobility networks and displacement distributions for the US and Italy. The baseline mobility networks captured for the US using **a** Cuebiq mobility insight data, **b** Safegraph, or **c** official census commuting flows display radically different structures. Whereas in the Cuebiq data, the data mostly comprises short-mid distance displacement characteristics of daily mobility, both Safegraph and the Census data include a large fraction of long trips that can be associated with the movements between different regions. The baseline mobility network for Italy using Cuebiq HDR data (**d**), Facebook data (**e**), or official census commuting flows (**f**) also displays clear differences: the Census data includes a larger fraction of short trips while the network described by Facebook has a significantly smaller number of connections. In the bottom panels, we illustrate the Complementary Cumulative Density Function (CCDF) of the displacement distribution $P(L)$ for the US (**g**) and Italy (**h**), computed for the datasets discussed above, and compare them with public flows derived from Google Location History data[20] and to what is observed for the US using mobile phone CDR[1]. In Italy, we use measures provided by Octotelematics' private vehicle's GPS blackboxes[4]. Since these datasets have different granularity, all CCDF have been computed only for $L > 40$ km. The resulting distributions display a huge variability between datasets, with Cuebiq and Mobile data characterized by a shorter tail in the US data, while Google data displays a curve suggesting the presence of two separate scales. In addition, Safegraph and the US census clearly display a totally different mobility scale. In Italy, the shorter tail is instead associated to Census and Octotelematics data, and Google, Cuebiq, and Facebook are displaying a similar long-tail trend.

group[6,15]. These differences reflected in the specific parameters of the functional forms taken by these distributions, that are typically analyzed in aggregated form (see refs. 4,20,32 and references therein).

In Fig. 1, we show the mobility networks and displacement distributions for the US and Italy computed using the different datasets described above. In panels a–f, we display the networks obtained for different datasets describing the baseline mobility across US and Italy. Due to the processes used for creating these datasets, the mobility networks are characterized by very different edge densities (see Table 1 in "Methods") which is visually manifest here (additional measures and details can be found in Supplementary Fig. 1). We then compute the complementary cumulative (CCDF) of the displacement distribution $P(L)$ (see Fig. 1g, h). This distribution has the advantage of capturing simultaneously information about the distances covered and the associated flows.

The CCDF of $P(L)$ is evaluated by progressively summing up the fraction of flows up to a distance $\leq L$. Depending on the type of datasets, these summations are made either by counting the number of individual displacements shorter than L or by summing up flows passing through edges connecting distances shorter than L. The latter is the case of all the data analyzed here, with the exception of the curves computed in other papers directly from mobile phone CDR[1] or GPS blackboxes[4].

These CCDFs, together with data coming from other aforementioned datasets, are compared in panels Fig. 1g, h, where we observe very large deviation depending on the dataset and the processing methodology (see the Supplementary Note for more details) used for analyzing the mobility of the same country. Similar comparisons are made in Supplementary Fig. 2 for France, Spain, and Portugal using Facebook and Google data along with mobile phone data coming from ref. 24). The observed differences appear larger than the changes associated with the changes in mobility due to lockdowns in the US and Italy in 2020 (see Supplementary Fig. 3), which represented, in some sense, a huge natural experiment and which we could safely consider as 'major disruptions'. We fit the curves for $P(L)$ to compare Google and Facebook data across 58 countries and used different functional forms for the fit (see "Methods" and Supplementary Fig. 4 where we observe, again, large differences between the curves observed in different datasets). We will discuss here the case of the truncated power law (TPL) form with two parameters. In Fig. 2a, we display these fitting parameters for 143 Google curves, 88 Facebook curves, the curves described in Fig. 1, and the time dependent curves of Supplementary Fig. 3. We see that the parameters of the TPL (but also for the other fitting models, see Supplementary Fig. 5) are distributed along a relatively narrow curve (that can be approximated by a sigmoidal function) where the curve exponent is a function of the scale. These results demonstrate the very large variability of mobility networks over different countries and data providers, pointing out differences that might bias models which are known to be sensitive to mobility, and in particular to long range mobility where the differences are more manifest, such as epidemics spreading. Additionally, they point to limitations of the different data design and the need for enhancing data collaborations to ensure the reproducibility of scientific studies based on mobility networks data.

Furthermore, the observed correlation between scale and exponent points to the existence an underlying lower dimensionality model. In Fig. 2b, we illustrate the set of TPL PDFs restricted to the condition set by the fitted relationship between scale and exponent. These curves present a growing part at small distances for the smaller scales—where the exponent is negative—and progressively converge towards a power law with exponent ≈1.5 for larger scales.

Lastly, an international comparison based on Facebook and Google data across 61 countries reveals that the statistical characteristics of human displacement differs heavily among different countries and datasets as significantly different tail behaviors are observed in the smaller and larger countries (see Supplementary Fig. 6). There is evidence[4]—coming from a single but high resolution dataset where the data processing methodology was fully in the hands of the authors—that the underlying physical process driving individual displacements should be unique, shaped by the hierarchical nature of transportation systems, and should follow the same logic everywhere. However, bounding human mobility studies to the geographical

Table 1 | Baseline networks characteristic numbers

Dataset	*N*	*E*	⟨*k*⟩	*D*
ITA Cuebiq	107/107	3735	34.9	0.329
ITA Facebook	110/110	926	8.4	0.077
ITA Census	110/110	4019	36.5	0.335
USA Cuebiq	3077/3108	60,181	19.6	0.006
USA Safegraph	3108/3108	1,574,604	506.6	0.163
USA Census	3108/3108	131,391	42.3	0.014

We display here the dimensions of the networks shown in Fig. 1. We compute the number of nodes N, the number of directed Edges E, the average (in- or out-) degree $\langle k\rangle = E/N$ and the edge density $D = E/(N(N-1))$. The Italian Cuebiq (ITA Cuebiq) network has been aggregated over different provinces which contains different numbers of nodes. Having neglected self-loops, the USA Cuebiq dataset has 31 disconnected nodes.

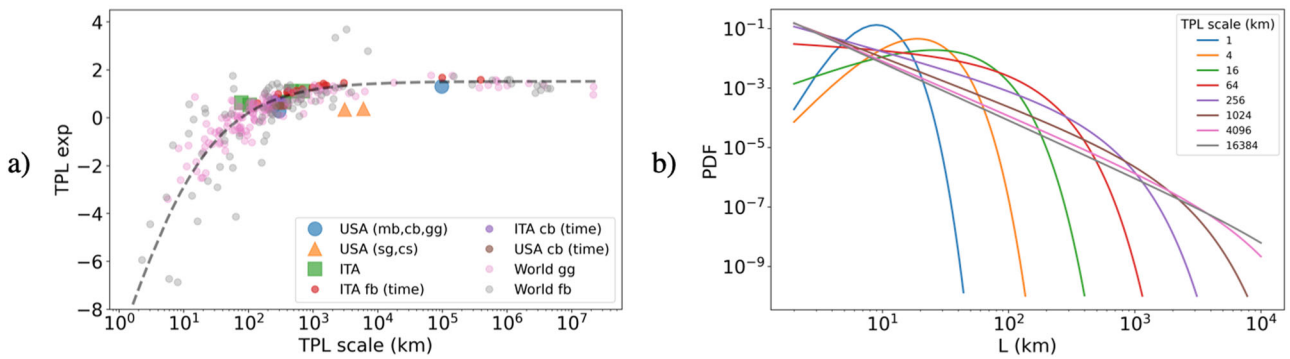

Fig. 2 | Comparison of the displacement distributions across datasets. In (**a**), we estimate the fitting parameters of the PDF (probability density function) $P(L)$ (scale and exponent) of various countries using a truncated power law. We observe a wide variance of these parameters. We include the results for the Fig. 1 for USA—Cuebiq (cb), Mobile (mb)[1], Google (gg), Safegraph (sg), Census (cs)—and Italy (as illustrated in Fig. 1 it also includes Octotelematics and Facebook (fb) data), the values for the longitudinal weekly analysis of Cuebiq and Facebook data, and the values estimated for the Google and Facebook baseline. We observe that all data points fall along a single curve, which suggests strong correlation between the free parameter and, ultimately, the existence of an underlying law having a single degree of freedom. This curve (dashed line) is estimated with a fit of the form $y = \frac{A}{1+x^{-1/B}} - C$ where $A \approx 22 \pm 1$, $B \approx 1.7 \pm 0.1$, and $C \approx 20 \pm 1$. We note how two USA data points (sg, cs) seem to deviate strongly from the average curve. This is because as they are capturing a behavior totally different from the other sources of data. In (**b**), we illustrate how this fitted relationship between exponent and scale translates into a family of Truncated Power Law PDFs (probability density functions) curves, where negative exponents illustrated in (**a**) are associated to growing trends for smaller distances for smaller scales, while the overall shape progressively converge towards a Power Law for larger scales.

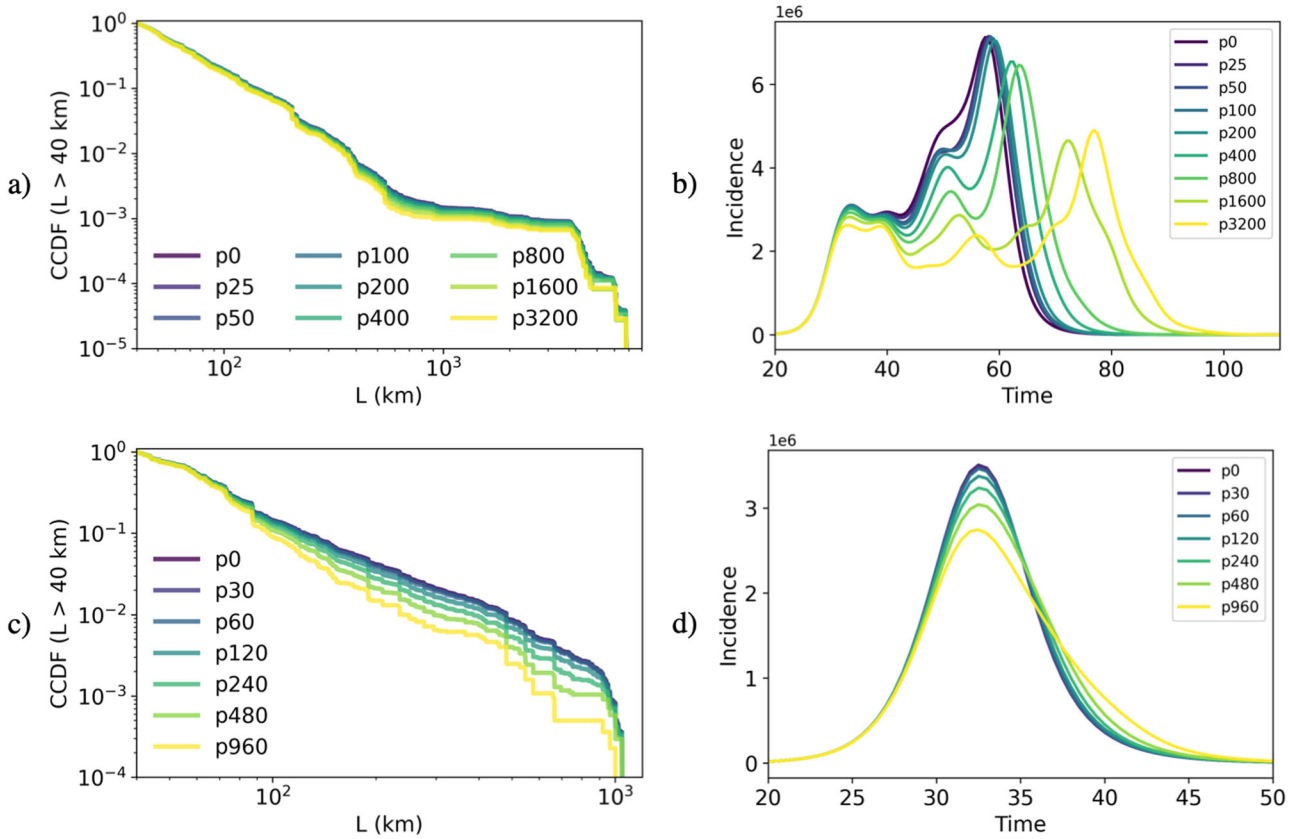

Fig. 3 | Effect of flow pruning on the spreading dynamics. In these figures, the notation p_x indicates the edge pruning with flow smaller than x. In (**a**) and (**c**), we show by examining the CCDF (Complementary Cumulative Distribution Function) of the displacement distribution that the pruning of flows does not affect significantly $P(L)$. Despite the small effect on the displacement distribution, we observe very different dynamics for disease spread on the corresponding mobility networks, in particular in the US case (**b**), while in the Italian case (**d**), the impact on the dynamics is marginal, probably due to the smaller size of this country.

constraint of national borders also influences the aggregated $P(L)$ as shown in Supplementary Fig. 6.

This effect, added to the data biases coming from both the socio-economical characteristics of behavior captured by the data, and the choices made in the processing of these data, renders the characterization of individual mobility from the distribution $P(L)$ of a single country not trivial.

The large sensitivity of spreading processes

The long range mobility information captured in individual trajectories datasets are used for a variety of applications, such as seasonal changes in population distribution[33], migration[34], tourism[35], and international mobility [36]. As a representative case study, we focus our attention on a class of models where human mobility plays an essential role: the spreading of an infectious disease. To better understand the consequences of the aforementioned variations observed on $P(L)$, we consider a standard metapopulation SIR model (see ref. 37 and the "Methods" section), where spatial patches play the role of nodes and human flows connect them. The choice of this process is justified by the high societal impact it has, since human mobility is a standard target for non-pharmaceutical interventions—such as curfews and lockdowns—to mitigate the evolution of infectious disease epidemics[18,19,38,39]. Also, it is worth remarking that this class of models is emblematic and widely adopted for practical applications: in principle, other models could be considered as well, but they might depend on a larger number of parameters and hyperparameters, thus making more difficult the analysis of the impact of mobility data on the final results. The choice of a metapopulation SIR keeps the number of parameters small while still providing a valuable model used for realistic infectious diseases.

We start with the Facebook data for 71 countries where at least 10 nodes are represented in the network. We first observe (see Supplementary Fig. 7a) that the average degree of the networks appear to be a (sub-linearly) growing function of the number of users, which means that more populated countries have denser networks. This is most likely due to a pruning procedure that removes edges with smaller flows for both privacy enhancing and disk space reduction. Networks with higher average degree have a larger fraction of shortcut nodes (as can be seen in Supplementary Fig. 7b, where shortcuts are defined as links at least twice as long as the average distance between a node and its closest neighbor). This biased pruning heavily influences the flows, and imposes in particular strong cutoffs on the longer connections (see more discussions about this point in Supplementary Fig. 8, where we use a toy model in order to analyze the relation between the density of edges and the displacement distribution). As a consequence of tampering the displacement tails, the spreading behavior on mobility networks ranges from lattice-like to a small-world behavior, depending on the dataset processing details. Also, with less shortcuts, the spreading dynamics is heavily penalized in networks captured with smaller user-bases and smaller average degrees as it is indeed observed in simulations (see again Supplementary Fig. 7).

The differences between the spreading dynamics in the US and Italy computed with different datasets are shown in Fig. 3. We first observe that for the US Safegraph and Census data, the results are exactly the same (figures shown in Supplementary Fig. 9). In contrast, simulations done with the Cuebiq data, being at a shorter scale, produce a more complex spreading behavior with two peaks. In Supplementary Fig. 10, we show how this first peak is characterized by a localized behavior reflecting a 'lattice like' spreading[40]. Italian data displays less differences, with the peak that appears not delayed but rather lowered for shorter scale Facebook and Census data.

To further investigate the differences observed between the US and Italy, we use the Cuebiq dataset and create a set of networks by pruning flows

below an increasing threshold. In Fig. 3a, c, we show that the displacement distribution remains unchanged. However, despite the small effect on $P(L)$, the pruning has a major impact on the spreading dynamics (Fig. 3b, d) that appears to be progressively delayed as the pruning is increased[40]. This is particularly true for the US case, while the Italian case—being at a smaller spatial scale—is not as sensitive to the pruning which induces a drop in the peak but not a delay. The differences observed in Fig. 3 as a consequence of the pruning are, again, smaller with respect to the differences observed when comparing across datasets and countries. This fact can be better appreciated in Supplementary Fig. 11, showing how the relative incidence at peak simulated with the SIR model are determined by the characteristics of the $P(L)$.

Discussion

While in general aware that data gathering and processing techniques used for human mobility are different and numerous technical limitations are present (see also the Supplementary Note for an overview of some of these technical limitations), the human mobility research community has often focused more on the opportunities associated with new datasets available rather than their limitations, under the optimistic assumption that the effect sizes of these biases would not strongly influence the final results.

In this paper, we have instead clearly illustrated the extremely wide differences between datasets of human movements which are commonly used interchangeably as a proxy for human behavior. Most studies in the field are based on the use of a single dataset describing mobility in a particular country and time and derived from a single data provider (see Supplementary Table II and the references therein indicated). In several cases, the data were prepared and released by a team different by those who ultimately analyzed them, not allowing to the researcher a complete control over their research methodology. Our paper highlights the limitation of this approach. On the one hand, our empirical findings from the analysis of multiple datasets call for a more transparent access to data and its protocols. On the other hand, our findings demonstrate that many of the limitations that were only supposed to slightly affect results do, in practice, significantly affect the insights about human behavior that can be obtained from mobility data.

Specifically, our study demonstrates that it is not sufficient to analyze human mobility and its multifaceted aspects by means of a single dataset: ensembles are needed to allow for a careful evaluation of potential biases in data gathering and processing protocols. The first step is to become more aware than before of such a critical impact: in particular, we noticed a series of technical aspects about these various datasets, that are often overlooked by both their producers and by scientists using them (see the SI for more details). In this respect, our analysis provides an overview of the limitations inherent to empirical studies and point towards the urgent need for setting shared standards in the design and elaboration of human mobility datasets. The focus was recently on how to model complex socio-technical systems[41], but their complexity also affects their empirical analysis, implying a special care is needed for their preparation due to their high sensitivity to specific processing.

Mobility datasets being passively collected from ICT data records, are necessarily influenced by the behavior that produces the record itself[42–45]. The method used is often undisclosed and results can be very sensitive to parametric choices. For instance, this is the case when aggregating data over any areal units which influences the observed results (see Supplementary Fig. 12), or when stop locations are defined according to radically different methodologies[46]. Also, pruning flows for anonymity reasons is a critical point and can have a dramatic large-scale impact as we saw for the Google or Facebook data. In particular, scale choice and pruning lead to different displacement distributions and affects particularly its tail behavior. Spreading processes over wide areas is more than the sum of the spreading processes over its parts and the tail of the displacement distribution is critical here, as we have shown with our application to epidemic dynamics.

The differences in methodologies might render meta-analysis and comparing different countries difficult, if not impossible. In the worst case, results obtained for a specific region might not be straightforwardly ported to another geographic area. Without verifying that such results are robust to changes in the protocols for data processing, the knowledge gained from one place—under the implicit assumption of dataset independence which we have only partially reproduced with our extensive analysis—could not agree with the one gained from another place, jeopardizing global policies.

Most of the problems we have outlined in this study could, however, be solved and we propose a series of good practices that largely align with what already largely discussed in the Mobile Phone Data Community[44,45]. Many researchers do follow good practices such as considering multiple datasets when possible and performing extensive robustness checks, against different preprocessing choices, and a few contributions[6,15,47,48] indeed already tested, or controlled for, some of the issues we illustrate in the Supplementary Note, but this kind of testing is, however, often unfortunately neglected or impossible. Certainly, encouraging a direct access to device-level data, although naturally strictly regulated to the higher anonymization and privacy-preserving standards, as Cuebiq and Mobile phone providers did, will enforce a high level data quality. Also, more transparency in the data production and pre-elaborating baselines and metadata (as Facebook did) is needed. Building of communities as encouraged by Cuebiq and Safegraph, creating open platforms and a share framework for mobility data[49] as proposed by initiatives such as DataCollaboratives[50] (especially associated to emergencies like COVID) are also measures that will improve the quality of data. Projects producing mobility datasets from CDRs using a shared and transparent methodology across several countries, such as those by Flowminder[51] or the Multi-MNO project[52] already represent excellent practical solutions in this sense.

On a more general perspective, it is always desirable that the study of complex systems relies on more than a single instance[53]. Even trivial operations on the data, like thresholding, can be very delicate[54] and require a robustness analysis that, either for lack of generalized access to multiple datasets or other plausible reasons, too often is overlooked. In fact, results might be platform-dependent[55] and data quantity cannot compensate for lack of data quality [56]. Therefore, a quantitative approach to social sciences aiming at the highest standards of reproducibility should go beyond single datasets, requiring by design to be tested with comprehensive meta-analyses[57].

Methods

Datasets

Census data. Census data provide statistical information about the urban and inter-urban commuters' flows using a widely statistical representative sample. The data can in principle be collected at the street address level of granularity, but is often aggregated at different administrative levels (district, municipality, provinces, counties, …). Census data typically covers the whole adult population in a country. However, the quality of these data as proxies for mobility may be limited by the fact that the home (or work) location in the national registers needs to be regularly updated as individuals move to new homes and are still registered at their original homeplace. This mismatch may vary strongly between countries as it depends on how enforced or rewarded is the update of the public register. Moreover, these data fail to capture other types of mobility besides home-work commuting of adults, such as the movements of tourists seasonal or freight traffic. At a urban level, census survey data are reasonably similar to the estimates coming from Mobile Phone or Twitter data[31] and allow for the opportunity of studying epidemic spreading considering the movements of different age groups[58,59], although the use at a national scale[59,60] can be influenced by the aforementioned limitations, as discussed in this paper. Here, we collected and analyzed official 2015 census data describing aggregated commuting patterns for the US[22], at the county level, and the official 2011 Census data for Italy[23], aggregated at the province level.

Mobile phone data. Call detail records (CDR) from mobile phone data were at the root of the first ICT data-informed attempts at modeling human mobility[1,2]. The information provided by the CDR are records of when and where a user exchanges information (audio, text, or data) with

an antenna. Unless some triangulation methods are applied, the spatial distancing between the antennas represent the characteristic granularity of the mobile phones trajectory data. In the earlier days of mobile communication, when it was limited to calls and SMS, the data temporal sampling was uneven and sparse, which would cause biases in how the trajectories were reconstructed[46], with an over-representation of long trips. This is most likely not the case with modern data, since smartphones are continuously connected to the mobile network for data exchange. Thanks to its wide coverage and the fact that it includes a wide range of mobility behaviors, mobile phone data represent a key asset that can be used for studying human interaction[61] and as proxies of human mobility for epidemic modeling[24,62] with applications in the study of the spread of a wide range of diseases[18,63–68]. The access to these data is usually limited to a small number of research groups that build tight partnerships with the mobile industry (with the exception of D4D challenges[69,70]), but the active collaboration between industry and academia in principle grants that the data processing can be custom built for the scientific use and the underlying assumption shared for reproducibility purposes. Here, we will reconstruct the displacement distribution function for the US on the basis of published fitting parameters[1], and those for Portugal, Spain, and France, from the data published for reproducibility purposes[24].

GPS data. Small blackboxes equipped with GPS trackers and accelerometers are installed in private vehicles for insurance reasons and record trajectory data. This type of data is naturally limited to a single transportation mode, but is able to capture locations with a great spatial accuracy (typically of order 10 m), and the dynamics of the trip with stops and velocity, which allows the reconstruction with high accuracy of a driver's displacements. The scientific use of a dataset describing the vehicular movements in Italy, provided by Octotelematics, has given the opportunity for exploring different facts of human mobility behavior[3,4,71–74]. However, with the exception of a small dataset provided for a data challenge[75], their access has been limited to a limited number of groups who had granted access to the raw data by an industrial partner. Here, we analyze the displacement statistics of Italian drivers in various cities as described in ref. 4.

Google datasets. Google data are elaborated from the opt-in service 'Google location history' and the trajectory data, capturing the movement of Android smartphone users, has been segmented using a machine learning technique[76] into a series of displacements between pairs of stop locations. The displacement have been then aggregated over a grid and on a weekly bases. The grid size is been described as a 5 km × 5 km grid in ref. 20 and as a 5 km^2 cell grid in other publications on similar datasets. To ensure the users' privacy, a network pruning has been applied where flows between cells is smaller than 100 users/week. The strength of this dataset is its wide international coverage. However, since longer connections have smaller flows, the pruning procedure systematically cuts inter-urban mobility and is a clear limitation of these datasets. In particular, this leads us to think that some results derived from Google data using this procedure are biased towards an over-representation of short movements. While the analysis at a urban scale[13,77–79] is probably marginally affected by this problem, the relative weight of long range connections is certainly biased. This bias might be one of the factors leading to the observation[20] of a steeper decline in the displacement distribution in less developed countries where, coincidentally, smartphones experience also a smaller market penetration and thus the effect of the threshold-based pruning of small long range flows can be naturally stronger. As we will see (see text and Fig. 3), this pruning effect also influences all analysis describing the disease spreading process over the mobility network at a country scale, such as those discussed in refs. 11,80. The Google data analyzed here has been shared through a Data4Good program. We do not, however, have access to the data and a reproducibility request associated to two published papers[11,20], was made in October 2020 is still pending at the time of the submission of this paper. We will therefore use the displacement statistics released by the authors of ref. 20 and which have been computed using bins of 1 km. The data covers almost all countries for any displacement larger than 1 km but we restricted our analysis to displacement larger than 5 km taking into account the cut-off induced by the underlying cell dimension.

Facebook. Facebook data is obtained by tracking the movements of the mobile phone application users that opted-in to the Location History and Background Location collection services[26]. The access is granted through a Data4Good platform which is accompanied by a limited description of the dataset and very limited methodological description. These datasets provided by the Facebook Data4Good can describe, in different forms, the mobility at different scales in correspondence to natural disasters (disaster maps) or the COVID-19 epidemics. The COVID-19 DiseaseMaps dataset included three descriptors of human mobility that have been used for research: (i) small scale mobility flows between small tiles of size ~1 km × 1 km (the dimension of the tiles varies between datasets[81]) and internal to administrative areas[12,16,81]; (ii) long range mobility flows between administrative areas[10,82]; (iii) co-location matrices computed at an administrative level[26,49,83]. Co-location matrices being intrinsically different from OD matrices, are not considered here. Similarly to what is happening to Google datasets, Facebook performs the pruning of small flows and long distances, thus introducing a strong bias and might severely affect results at a large scale (and less at a urban level), and leading to national origin-destination matrices that are visibly sparse as observed in ref. 10, Fig. 1e and Supplementary Fig. 1 (This effect of pruning appears as reduced in co-location matrices.). Also, we note that Facebook also provides 'baseline' values that represent a typical flow prior to the event (disaster or disease) characterizing the dataset. The strength of the Facebook data is its wide international coverage, coupled with the possibility of providing metadata such as user gender. Limitations here come from the data processing under the form of small flow pruning and the fragmentation of the network into smaller sub-networks. For instance, U.S.A. flows for the COVID-19 disease maps are provided only at state level, with no information about the movements between states. Here, we will use the mobility network where the nodes represent administratives areas and the edges the long range mobility flows between these areas. The flows are aggregated every 8 h and we obtain temporal networks. If the flow is below a certain threshold (that we estimate it to be of about 10 users), it is not reported in the data. For these temporal networks, we aggregate flows on a weekly basis. Similarly, we also define an aggregated baseline network combining all baseline flows reported for different 8 h bins and day of the week. Baseline flows are present, however, only for origin-destination pairs and time bins where it the flow reported is above the filtering threshold in at least one of the weeks comprised in the dataset, thus some information is necessarily lost also in this case.

For epidemic simulations, we also used, as a proxy for population, the baseline values describing the number of Facebook active users recorded in a given node[81]. Similarly to flows, active users are reported in bins of 8 h and for each day of the week and have been accordingly averaged over a whole week. This proxy has been shown to be effective for this type of studies (see also Supplementary Fig. 14). Derived measures depending on this number, such as the social connectedness index between two spatial patches, have been used to model sub-national trade[84] and capital[85]. The same data has been shown to provide a good proxy to reproduce the geographic spread of COVID-19[86] and compliance with non-pharmaceutical interventions (such as mobility restrictions) during the pandemic[87] to mention a few emblematic successful applications.

Cuebiq. Cuebiq collect mobility data of anonymized mobile app users who opted-in to a large number of different location-based services in different countries including U.S.A., U.K., Italy, Spain, France and Germany. Access to this data has been provided under data governance frameworks to address the COVID-19 emergency via a Data4Good program that constitute a clear attempt at building a collaborative

environment around the data. Cuebiq provided to the members of the program access to device-level privacy-enhanced data, where noise is added to home and work locations at the census block group level, and stops associated with privacy-sensitive locations are removed entirely from the dataset, in order to preserve privacy. This data has been used in the early days of the pandemic to produce mobility reports[28,88,89], often accompanied by interactive visualization of the mobility reduction patterns. A clear advantage of these datasets is the access to the device-level trajectory data at GPS precision, from which it has been possible to study the behavioral changes associated to lockdowns[90–92]. The limitation coming with this abundance of data is the need for building a strong preprocessing pipeline in order to analyze this data, but nevertheless allows for a tailored and transparent design of the segmentation algorithm. Here, we use Cuebiq-HDR mobility data for Italy derived from the device-level data and aggregated weekly at the province scale by Pepe and collaborators[28]. We also use flows describing the movements across the U.S.A., computed directly by Cuebiq and provided within the framework of the Data4Good program. The flows are here also aggregated weekly and at the county scale, based on a proprietary segmentation method. We computed the baseline flows aggregating those of the weeks preceding the beginning of lockdowns in Italy and U.S.A., respectively, rescaled the flows to the total country population and used the associated census population for epidemic modeling.

Safegraph. Safegraph collects statistics about visiting patterns of different points of interests (PoI) by aggregating anonymised location data from mobile applications. The data covers the US only, where the home location is identified for each user at the level of census block group. Safegraph data activated a Data4Good program that gives free access to Academics for non-commercial work. Safegraph actively attempted at building a community around their data by organizing regular seminars and a platform where results and issues can be exchanged. The limitation of this dataset is that, since it is focusing on PoIs, mobility flows are not directly available from the data[15], as the flows recorded are between home locations and visited PoI. This means that two subsequent visits to two location A and B would be recorded not as a movement between A and B but as two movements from Home to A and from Home to B. Deriving the mobility network at the urban level requires a rather complex procedure[15], which is likely not suitable for being extended at a country scale. Nevertheless, these mobility flows have been released as OD matrices to describe population flows during the COVID-19 pandemic[29] and used to reconstruct mobility at long range[93]. The advantages of this data is that there is a great abundance of detailed information about the users activity at PoI level, including social distancing estimates that have been also subject to scientific use[87,94]. Here, we will use the aggregated flows published in Nature Scientific Data[29], aggregated weekly and at county level, and used a 2-month period before the U.S.A. lockdown do define the baseline flows.

Network features. We recap in Table 1, the main features of the baseline networks obtained from the various datasets used for Italy and the US.

Curve fitting as analysis method

In this paper we use the distribution of displacements $P(L)$ as one of the analysis tools to inspect at the same time the flows and the distance covered in a flow network/origin destination matrix. In the past years, there has been an open discussion about the functional form of this distribution, that we know to be governed by the multi-scale characteristics of human mobility[4,5]. As in many similar cases, we use curve fitting as an analytic tool for extracting information from the data collected. Curve fitting is clearly a method limited by the functional form chosen. A large number of functional forms have been used for mobility datasets, and the most common ones are:

1. a truncated power law (TPL) with three[1] or two[19] free parameters;
2. a lognormal (LN)[32];
3. model-driven generalized gamma functions[4].

An exponential form is also sometimes considered at a shorter scale, but can be captured by the exponential cutoff using TPL and model-driven generalized gamma functions. In our analysis, we attempted fits using 2 and 3 parametric TPL, LN and three different model-driven forms: the Random Uncorrelated Accelerations (RUA, 1 parameter), the Random Acceleration Kicks (RAK, 2 parameters) and the Weibull (WB, 2 parameters).

The RUA follows the functional form $P(L) \propto L^{-\frac{3}{4}} \exp(-(L/L_s)^{\frac{1}{2}})$, and is what expected when a traveller speed follows a Brownian motion[4], the RAK is in practice a generalization of RUA, as follows the similar form $P(L) \propto L^{-\gamma} \exp(-(L/L_s)^{\frac{1}{2}})$, where γ is a free parameter, and represents the saddle points approximation for a model where a traveller accelerates and decelerates at fixed increments following a Poisson process[4].

The plots were technically carried out by fitting directly the CCDF function, without any data binning, using the python *lmfit* library, which performs a (non-linear) least square fit. This procedure clearly weights more the short distances in the distribution, where data is more dense, and basically ignores the tail. Since the detailed characterization of tail for the curves at hand was essential, we compensated by fitting the logarithm of the y-axis of the empirical curves. The results are satisfactory, as illustrated in Supplementary Fig. 13.

Metapopulation SIR model

In order to understand the geographical diffusion of diseases, one has to combine the microscopic contagion processes with the long-range disease propagation due to human mobility across different spatial scales. In order to tackle this problem, epidemic modeling has relied on reaction-diffusion dynamics in metapopulations[95]. Metapopulations can be thought as nodes of a complex network of spatial patches, where links encode human flows from one place to another and are responsible for between-patch transmission[96].

We denote by M the number of patches, N the total number of agents and N_i the population of the i-th patch. At any time, we have $\sum_i N_i = N$ and, if the system is closed (i.e., there are no births and deaths), this number N is conserved. The mobility of the agents between the patches is ruled by a weighted adjacency matrix $\mathbf{W}$, whose entry W_{ij} is the flux from patch i to patch j. The probability P_{ij} that an agent placed in i moves to j must be proportional to the flux W_{ij} and reads (as in ref. 95)

$$P_{ij} = \frac{W_{ij}}{\sum_{j=1}^{M} W_{ij}} \tag{1}$$

At this point one has to introduce the reaction dynamics, that takes place independently within the patches. A very wide-used and simple model for this is the so-called SIR-model: agents belongs either to the susceptible (S), the infected (I) or the recovered (R) compartment; therefore, in each patch i we have that $S_i + I_i + R_i = N_i$. The allowed reactions are the following: one agent that is in patch i can move from the S_i to the I_i state by getting the infection from another infected agent with infection rate β[95]

$$S_i + I_i \xrightarrow{\beta} I_i + I_i \tag{2}$$

and one agent (in patch i) can move from the I_i to the R_i state by healing from the infection with recovery rate μ[95]

$$I_i \to {}^{\mu} R_i \tag{3}$$

therefore, the continuous-time equations for the infection dynamic are

$$\begin{cases} \frac{dS_i}{dt} = -\beta I_i \frac{S_i}{N_i} \\ \frac{dI_i}{dt} = \beta I_i \frac{S_i}{N_i} - \mu I_i \\ \frac{dR_i}{dt} = \mu I_i \end{cases} \tag{4}$$

Now, if we work under the assumption that the state of an agent does not affect its diffusive behavior, the mobility for all the agents is described by a

unique mobility matrix **P**, and the continuous-time equation relative to the mobility for a generic compartment X is

$$\frac{dX_i}{dt} = \sum_{j=1}^{M} \mathcal{P}_{ji} X_j - \sum_{j=1}^{M} \mathcal{P}_{ij} X_i \tag{5}$$

Therefore by summing up Eqs. (4) and (5), one obtains the system of $3M$ differential equations of the model (that basically is the same as ref. 97 with $\epsilon = 1$)

$$\begin{cases} \frac{dS_i}{dt} = -\beta I_i \frac{S_i}{N_i} + \sum_{j=1}^{M} \mathcal{P}_{ji} S_j - \sum_{j=1}^{M} \mathcal{P}_{ij} S_i \\ \frac{dI_i}{dt} = \beta I_i \frac{S_i}{N_i} - \mu I_i + \sum_{j=1}^{M} \mathcal{P}_{ji} I_j - \sum_{j=1}^{M} \mathcal{P}_{ij} I_i \\ \frac{dR_i}{dt} = \mu I_i + \sum_{j=1}^{M} \mathcal{P}_{ji} R_j - \sum_{j=1}^{M} \mathcal{P}_{ij} R_i \end{cases} \tag{6}$$

This system of equations cannot be solved analytically but only numerically; notice that for the basic reproductive ratio we simply have $R_0 = \beta/\mu$ (see for example ref. 97 for details).

Simulation details

All the simulations were done by choosing the free parameters β and μ in order to have $R_0 = 2.6$, building the mobility matrix from the data and initializing the population of the patches with the data. The local population was set as proportional to the reported baseline userbase for Facebook data in countries different from Italy, while for Facebook Italy, Cuebiq, Safegraph, and Census data the resident population has been used. The other free parameter to be chosen (over the total duration of the simulation, that was selected to be $t = 200$ days just to be sure to reach the end of the epidemic) is the time step, that was chosen as $dt = 0.5$ days. The effect of this parametric choice has been tested by performing simulations with $dt = 0.1$ days, which yielded exactly the same numerical results. Given the initial conditions and the parameters, the code for the simulations solves Eq. (6) by using the built-in `ode` function of the software `R`, providing the time evolution for S, I, R individuals in all the patches at each time step starting from $t = 0$. All the simulations have at $t = 0$ a fully susceptible population but one infected (the seed), whose location was fixed, in each network, in the node with the larger population. From the time evolution, the code trivially calculates the total incidence (namely the total number of new infected) at each time step, the number of recovered at the end of the epidemic in each patch and the attack rate in each patch. Notice that by the moment that in some networks unconnected components are present, in order to calculate the attack rate the recovered were not divided by the total population of the network, but only by the population of the component of the network in which the initial seed was placed.

Data availability

All dataset used and their sources are listed in Supplementary Table II. Census (ITA[23], USA[22]), Cuebiq ITA[28], Gonzalez 2008[1], Google[20], Mobile[24] (FRA, ESP, POR) and Safegraph[29] data are available through the referred URLs or papers. Elaborated data derived from Cuebiq USA[27], Facebook[25], and Octotelematics[4] datasets are available from the authors upon reasonable request, while the access to the associated raw data is restricted as the data have been used under licence and are not publicly available.

Code availability

The code used for data processing and curve fitting are available upon request, to be sent to the first author.

Received: 17 November 2022; Accepted: 9 December 2024;
Published online: 24 December 2024

References

1. Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 779–782 (2008).
2. Song, C., Koren, T., Wang, P. & Barabási, A.-L. Modelling the scaling properties of human mobility. *Nat. Phys.* **6**, 818–823 (2010).
3. Pappalardo, L. et al. Returners and explorers dichotomy in human mobility. *Nat. Commun.* **6**, 1–8 (2015).
4. Gallotti, R., Bazzani, A., Rambaldi, S. & Barthelemy, M. A stochastic model of randomly accelerated walkers for human mobility. *Nat. Commun.* **7**, 1–7 (2016).
5. Alessandretti, L., Aslak, U. & Lehmann, S. The scales of human mobility. *Nature* **587**, 402–407 (2020).
6. Moro, E., Calacci, D., Dong, X. & Pentland, A. Mobility patterns are associated with experienced income segregation in large US cities. *Nat. Commun.* **12**, 1–10 (2021).
7. Schläpfer, M. et al. The universal visitation law of human mobility. *Nature* **593**, 522–527 (2021).
8. Buckee, C. O. et al. Aggregated mobility data could help fight COVID-19. *Science* **368**, 145–146 (2020).
9. Nanni, M. et al. Give more data, awareness and control to individual citizens, and they will help COVID-19 containment. *Ethics Inf. Technol.* **23**, 1–6 (2021).
10. Bonaccorsi, G. et al. Economic and social consequences of human mobility restrictions under COVID-19. *Proc. Natl Acad. Sci. USA* **117**, 15530–15535 (2020).
11. Rader, B. et al. Crowding and the shape of COVID-19 epidemics. *Nat. Med.* **26**, 1829–1834 (2020).
12. Kissler, S. M. et al. Reductions in commuting mobility correlate with geographic differences in SARS-CoV-2 prevalence in New York City. *Nat. Commun.* **11**, 1–6 (2020).
13. Hazarie, S., Soriano-Paños, D., Arenas, A., Gómez-Gardeñes, J. & Ghoshal, G. Interplay between population density and mobility in determining the spread of epidemics in cities. *Commun. Phys.* **4**, 1–10 (2021).
14. Lemey, P. et al. Untangling introductions and persistence in COVID-19 resurgence in Europe. *Nature* **595**, 713–717 (2021).
15. Chang, S. et al. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* **589**, 82–87 (2021).
16. Mena, G. E. et al. Socioeconomic status determines COVID-19 incidence and related mortality in Santiago, Chile. *Science* **372**, eabg5298 (2021).
17. Mazzoli, M. et al. Interplay between mobility, multi-seeding and lockdowns shapes COVID-19 local impact. *PLoS Comput. Biol.* **17**, e1009326 (2021).
18. Schlosser, F. et al. COVID-19 lockdown induces disease-mitigating structural changes in mobility networks. *Proc. Natl Acad. Sci. USA* **117**, 32883–32890 (2020).
19. Kraemer, M. U. et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493–497 (2020).
20. Kraemer, M. U. G. et al. Mapping global variation in human mobility. *Nat. Hum. Behav.* **4**, 800–810 (2020).
21. Ruktanonchai, N. W. et al. Assessing the impact of coordinated COVID-19 exit strategies across Europe. *Science* **369**, 1465–1470 (2020).
22. US Census 2011-2015 5-Year ACS Commuting Flows. https://www.census.gov/data/tables/2015/demo/metro-micro/commuting-flows-2015.html (2015). (Accessed on December 2020).
23. ISTAT Matrici di contiguità, distanza e pendolarismo. https://www.istat.it/it/archivio/157423 (2011). (Accessed on December 2020).
24. Tizzoni, M. et al. On the use of human mobility proxies for modeling epidemics. *PLoS Comput. Biol.* **10**, e1003716 (2014).
25. Facebook Data For Good. https://dataforgood.facebook.com/dfg/tools (2020). (Accessed on November 2021).

26. Iyer, S. et al. Large-scale measurement of aggregate human colocation patterns for epidemiological modeling. *Epidemics* **42**, 100663 (2023).
27. Cuebiq Data For Good. https://www.cuebiq.com/about/data-for-good (2020). (Accessed on November 2021).
28. Pepe, E. et al. COVID-19 outbreak response, a dataset to assess mobility changes in Italy following national lockdown. *Sci. Data* **7**, 1–7 (2020).
29. Kang, Y. et al. Multiscale dynamic human mobility flow dataset in the us during the COVID-19 epidemic. *Sci. Data* **7**, 1–13 (2020).
30. Bazzani, A., Giorgini, B., Rambaldi, S., Gallotti, R. & Giovannini, L. Statistical laws in urban mobility from microscopic GPS data in the area of Florence. *J. Stat. Mech.: Theory Exp.* **2010**, P05001 (2010).
31. Lenormand, M. et al. Cross-checking different sources of mobility information. *PLoS ONE* **9**, e105184 (2014).
32. Alessandretti, L., Sapiezynski, P., Lehmann, S. & Baronchelli, A. Multi-scale spatio-temporal analysis of human mobility. *PLoS ONE* **12**, e0171686 (2017).
33. Deville, P. et al. Dynamic population mapping using mobile phone data. *Proc. Natl Acad. Sci. USA* **111**, 15888–15893 (2014).
34. Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484**, 5 (2012).
35. Lenormand, M., Gonçalves, B., Tugores, A. & Ramasco, J. J. Human diffusion and city influence. *J. R. Soc. Interface* **12**, 20150473 (2015).
36. Hawelka, B. et al. Geo-located Twitter as proxy for global mobility patterns. *Cartogr. Geogr. Inf. Sci.* **41**, 260–271 (2014).
37. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic processes in complex networks. *Rev. Mod. Phys.* **87**, 925 (2015).
38. Maier, B. F. & Brockmann, D. Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. *Science* **368**, 742–746 (2020).
39. Davis, J. T. et al. Cryptic transmission of SARS-CoV-2 and the first COVID-19 wave. *Nature* **600**, 127–132 (2021).
40. Gross, B. & Havlin, S. Epidemic spreading and control strategies in spatial modular network. *Appl. Netw. Sci.* **5**, 1–14 (2020).
41. Citron, D. T. et al. Comparing metapopulation dynamics of infectious diseases under different models of human movement. *Proc. Natl Acad. Sci. USA* **118**, e2007488118 (2021).
42. Bonnel, P., Hombourger, E., Olteanu-Raimond, A.-M. & Smoreda, Z. Passive mobile phone dataset to construct origin-destination matrix: potentials and limitations. *Transp. Res. Proc.* **11**, 381–398 (2015).
43. Olteanu, A., Castillo, C., Diaz, F. & Kıcıman, E. Social data: biases, methodological pitfalls, and ethical boundaries. *Front. Big Data* **2**, 13 (2019).
44. ESCAP, U. *Handbook on the Use of Mobile Phone Data for Official Statistics* (2019).
45. De Broe, S. et al. Updating the paradigm of official statistics: new quality criteria for integrating new data and methods in official statistics. *Stat. J. IAOS* **37**, 343–360 (2021).
46. Gallotti, R., Louf, R., Luck, J.-M. & Barthelemy, M. Tracking random walks. *J. R. Soc. Interface* **15**, 20170776 (2018).
47. Schlosser, F., Sekara, V., Brockmann, D. & Garcia-Herranz, M. Biases in human mobility data impact epidemic modeling. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2112.12521 (2021).
48. Mercier, A., Scarpino, S. & Moore, C. Effective resistance against pandemics: mobility network sparsification for high-fidelity epidemic simulations. *PLOS Comput. Biol.* **18**, e1010650 (2022).
49. Kishore, N. et al. Measuring mobility to monitor travel and physical distancing interventions: a common framework for mobile phone data analysis. *Lancet Digit. Health* **2**, e622–e628 (2020).
50. Data Collaboratives. https://datacollaboratives.org (2011). (Accessed on October 2022).
51. Flowminder standards in producing mobility and population estimates from call details records in low- and middle-income countries. https://www.flowminder.org/resources/publications-reports/flowminder-standards-in-producing-mobility-and-population-estimates-from-call-details-records-in-low-and-middle-income-countries (2023). (Accessed on October 2023).
52. Reusing mobile network operator data for official statistics: the case for a common methodological framework for the European Statistical System – 2023 edition. https://ec.europa.eu/eurostat/en/web/products-statistical-reports/w/ks-ft-23-001 (2023). (Accessed on October 2023).
53. Parisi, G. Complex systems: a physicist's viewpoint. *Phys. A: Stat. Mech. Appl.* **263**, 557–564 (1999).
54. Cantwell, G. T. et al. Thresholding normally distributed data creates complex networks. *Phys. Rev. E* **101**, 062302 (2020).
55. Malik, M. M. & Pfeffer, J. Identifying platform effects in social media data. In *Tenth International AAAI Conference on Web and Social Media* (2016).
56. Bradley, V. et al. Unrepresentative big surveys significantly overestimate US vaccine uptake. *Nature* **600**, 695–700 (2021).
57. Broido, A. D. & Clauset, A. Scale-free networks are rare. *Nat. Commun.* **10**, 1–10 (2019).
58. Dalziel, B. D., Pourbohloul, B. & Ellner, S. P. Human mobility patterns predict divergent epidemic dynamics among cities. *Proc. Roy. Soc. B: Biol. Sci.* **280**, 20130763 (2013).
59. Arenas, A. et al. Modeling the spatiotemporal epidemic spreading of COVID-19 and the impact of mobility and social distancing interventions. *Phys. Rev. X* **10**, 041055 (2020).
60. Gatto, M. et al. Spread and dynamics of the COVID-19 epidemic in Italy: effects of emergency containment measures. *Proc. Natl Acad. Sci. USA* **117**, 10484–10491 (2020).
61. Sobolevsky, S. et al. Delineating geographical regions with networks of human interactions in an extensive set of countries. *PLoS ONE* **8**, e81707 (2023).
62. Panigutti, C., Tizzoni, M., Bajardi, P., Smoreda, Z. & Colizza, V. Assessing the use of mobile phone data to describe recurrent mobility patterns in spatial epidemic models. *R. Soc. Open Sci.* **4**, 160950 (2017).
63. Wesolowski, A. et al. Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proc. Natl Acad. Sci. USA* **112**, 11887–11892 (2015).
64. Ruktanonchai, N. W. et al. Identifying malaria transmission foci for elimination using human mobility data. *PLoS Comput. Biol.* **12**, e1004846 (2016).
65. Kramer, A. M. et al. Spatial spread of the West Africa Ebola epidemic. *R. Soc. Open Sci.* **3**, 160294 (2016).
66. Bosetti, P. et al. Heterogeneity in social and epidemiological factors determines the risk of measles outbreaks. *Proc. Natl Acad. Sci. USA* **117**, 30118–30125 (2020).
67. Pullano, G., Valdano, E., Scarpa, N., Rubrichi, S. & Colizza, V. Evaluating the effect of demographic factors, socioeconomic factors, and risk aversion on mobility during the COVID-19 epidemic in France under lockdown: a population-based study. *Lancet Digit. Health* **2**, e638–e649 (2020).
68. Valdano, E., Okano, J. T., Colizza, V., Mitonga, H. K. & Blower, S. Using mobile phone data to reveal risk flow networks underlying the HIV epidemic in Namibia. *Nat. Commun.* **12**, 2837–10 (2021).
69. Blondel, V. D. et al. Data for development: the d4d challenge on mobile phone data. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1210.0137 (2012).
70. de Montjoye, Y.-A., Smoreda, Z., Trinquart, R., Ziemlicki, C. & Blondel, V. D. D4D-Senegal: the second mobile phone data for development challenge. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1407.4885 (2014).
71. Gallotti, R., Bazzani, A. & Rambaldi, S. Towards a statistical physics of human mobility. *Int. J. Mod. Phys. C.* **23**, 1250061 (2012).
72. Pappalardo, L., Rinzivillo, S., Qu, Z., Pedreschi, D. & Giannotti, F. Understanding the patterns of car travel. *Eur. Phys. J. Spec. Top.* **215**, 61–73 (2013).
73. Gallotti, R., Bazzani, A., Degli Esposti, M. & Rambaldi, S. Entropic measures of individual mobility patterns. *J. Stat. Mech.: Theory Exp.* **2013**, P10022 (2013).

74. Gallotti, R., Bazzani, A. & Rambaldi, S. Understanding the variability of daily travel-time expenditures using GPS trajectory data. *EPJ Data Sci.* **4**, 1–14 (2015).
75. Barlacchi, G. et al. A multi-source dataset of urban life in the city of Milan and the province of Trentino. *Sci. Data* **2**, 1–15 (2015).
76. Kirmse, A., Udeshi, T., Bellver, P. & Shuma, J. Extracting patterns from location history. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 397–400 (2011).
77. Bassolas, A. et al. Hierarchical organization of urban mobility and its connection with city livability. *Nat. Commun.* **10**, 1–10 (2019).
78. Aguilar, J. et al. Impact of urban structure on infectious disease spreading. *Sci. Rep.* **12**, 3816 (2022).
79. Barbosa, H. et al. Uncovering the socioeconomic facets of human mobility. *Sci. Rep.* **11**, 1–13 (2021).
80. Ruktanonchai, C. W. et al. Practical geospatial and sociodemographic predictors of human mobility. *Sci. Rep.* **11**, 15389 (2021).
81. Kishore, N. et al. Lockdowns result in changes in human mobility which may impact the epidemiologic dynamics of sars-cov-2. *Sci. Rep.* **11**, 1–12 (2021).
82. Galeazzi, A. et al. Human mobility in response to COVID-19 in France, Italy and UK. *Sci. Rep.* **11**, 1–10 (2021).
83. Chang, M.-C. et al. Variation in human mobility and its impact on the risk of future COVID-19 outbreaks in Taiwan. *BMC Public Health* **21**, 1–10 (2021).
84. Bailey, M. et al. International trade and social connectedness. *J. Int. Econ.* **129**, 103418 (2021).
85. Kuchler, T., Li, Y., Peng, L., Stroebel, J. & Zhou, D. Social proximity to capital: implications for investors and firms. *Rev. Financ. Stud.* **35**, 2743–2789 (2022).
86. Kuchler, T., Russel, D. & Stroebel, J. Jue insight: the geographic spread of COVID-19 correlates with the structure of social networks as measured by Facebook. *J. Urban Econ.* **127**, 103314 (2022).
87. Charoenwong, B., Kwan, A. & Pursiainen, V. Social connections with COVID-19-affected areas increase compliance with mobility restrictions. *Sci. Adv.* **6**, eabc3054 (2020).
88. Santana, C. et al. COVID-19 is linked to changes in the time–space dimension of human mobility. *Nat. Hum. Behav.* **7**, 1729–1739 (2023).
89. Klein, B. et al. Reshaping a nation: mobility, commuting, and contact patterns during the COVID-19 outbreak. Northeastern University-Network Science Institute Report (2020).
90. Gauvin, L. et al. Socioeconomic determinants of mobility responses during the first wave of COVID-19 in Italy: from provinces to neighbourhoods. *J. R. Soc. Interface* **18**, 20210092 (2021).
91. Hunter, R. F. et al. Effect of COVID-19 response policies on walking behavior in us cities. *Nat. Commun.* **12**, 1–9 (2021).
92. Lucchini, L. et al. Living in a pandemic: changes in mobility routines, social activity and adherence to COVID-19 protective measures. *Sci. Rep.* **11**, 24452 (2021).
93. Hou, X. et al. Intracounty modeling of COVID-19 infection with human mobility: assessing spatial heterogeneity with business traffic, age, and race. *Proc. Natl Acad. Sci. USA* **118**, e2020524118 (2021).
94. Weill, J. A., Stigler, M., Deschenes, O. & Springborn, M. R. Social distancing responses to COVID-19 emergency declarations strongly differentiated by income. *Proc. Natl Acad. Sci. USA* **117**, 19658–19660 (2020).
95. Brockmann, D. & Helbing, D. The hidden geometry of complex, network-driven contagion phenomena. *Science* **342**, 1337–1342 (2013).
96. Hagenaars, T., Donnelly, C. & Ferguson, N. Spatial heterogeneity and the persistence of infectious diseases. *J. Theor. Biol.* **229**, 349–359 (2004).
97. Castioni, P., Gallotti, R. & De Domenico, M. Critical behavior in interdependent spatial spreading processes with distinct characteristic time scales. *Commun. Phys.* **4**, 1–10 (2021).

Acknowledgements

We thank Cuebiq and Facebook for providing us free access to their Flow Network Data through their Data for Good programs. R.G. acknowledges the support of the PNRR ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing (CN00000013), under the NRRP MUR program funded by the NextGenerationEU. MDD acknowledges partial financial support from the MUR - PNC (DD n. 1511 30-09-2022) Project no. PNC0000002, DigitAl lifelong pRevEntion (DARE), from the INFN grant “LINCOLN” and from MUR funding within the PRIN 2022 PNRR (DD n. 1214 31-07-2023) Project no. P2022A889F.

Author contributions

R.G. and M.D.D. designed the research and collected the data. D.M. and M.D.D. developed the spatial epidemic models. R.G. conducted the numerical data analysis and produced the figures. R.G., M.B., and M.D.D. analyzed and interpreted the results. All authors wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42005-024-01909-x.

Correspondence and requests for materials should be addressed to Riccardo Gallotti or Manlio De Domenico.

Peer review information *Communications Physics* thanks Roger Guimerá and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.