Gaps in mobility data and implications for modelling epidemic spread: a scoping review and simulation study

Jack Wardle1*, Sangeeta Bhatia1*, Moritz U G Kraemer2, Pierre Nouvellet1, 3, Anne Cori1

1 MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, UK.
2 Department of Zoology, University of Oxford, Oxford, UK.
3 School of Life Sciences, University of Sussex, Brighton, UK.

* equal contribution

Abstract

Reliable estimates of human mobility are important for understanding the spatial spread of infectious diseases and the effective targeting of control measures. However, when modelling infectious disease dynamics, data on human mobility at an appropriate temporal or spatial resolution are not always available, leading to the common use of model-derived mobility proxies. In this study we reviewed the different data sources and mobility models that have been used to characterise human movement in Africa. We then conducted a simulation study to better understand the implications of using human mobility proxies when predicting the spatial spread and dynamics of infectious diseases.

We found major gaps in the availability of empirical measures of human mobility in Africa, leading to mobility proxies being used in place of data. Empirical data on subnational mobility were only available for 17/54 countries, and in most instances, these data characterised long-term movement patterns, which were unsuitable for modelling the spread of pathogens with short generation times (time between infection of a case and their infector). Results from our simulation study demonstrated that using mobility proxies can have a substantial impact on the predicted epidemic dynamics, with complex and non-intuitive biases. In particular, the predicted times and order of epidemic invasion, and the time of epidemic peak in different locations can be underestimated or overestimated, depending on the types of proxies used and the country of interest.

Our work underscores the need for regularly updated empirical measures of population movement within and between countries to aid the effective and sustained prevention and control of infectious diseases.
Introduction

Human movement is a key determinant of the spatial spread of infectious diseases as evidenced by the spread of Ebola [1, 2], yellow fever [3, 4], and most recently, SARS-CoV-2 [5, 6]. The emergence and subsequent international spread of the latter and its variants of concern (VOCs) highlight the speed and scale at which infectious pathogens can spread around the globe [7–9].

Models of infectious disease dynamics that incorporate human movement have been used to generate insights into the spatial spread of pathogens. They can help identify locations that are most susceptible to receive imported cases [10–12], assess the effectiveness of travel restrictions [13, 14], and inform the allocation of scarce resources such as vaccines, drugs, personal protective equipment or specialist healthcare staff [15–18]. Spatial models of infectious disease spread typically rely on human population movement data. These can be empirical measurements, for example flight passenger statistics, information from national population censuses, or mobile phone call detail records [19, 20]. More recently, a few high resolution datasets based on mobile phone GPS data have been made available to researchers and public health officials and used to analyse within-country movement patterns during the COVID-19 pandemic [21–23].

However, empirical data on human movement are not routinely collected, particularly in low- and middle-income countries [24]. Moreover, where empirical data are available, these are not always at a spatial and temporal resolution that matches the scale and pace at which pathogens spread [25]. Although mobile phone usage is increasingly being used to understand movement of populations at a fine resolution, these data are known to be biased in systematic ways by limited smartphone ownership or uneven mobile phone coverage [26, 27], and typically do not capture international movement. They are also expensive and time-consuming to acquire, limiting their potential use in real-time analysis. Finally, privacy concerns constrain wider sharing of these data [28], preventing the reproducibility of analyses based on them.

In the absence of empirical data, infectious disease modelling studies often rely on human movement models such as gravity and radiation models [29–31]. These models are sometimes calibrated using data from nearby countries [4, 29], long-term movement data from Demographic and Health Surveys (DHS), or observed patterns of disease spread [32]. By definition, these mobility proxies cannot be validated against local human movement data which do not exist. The potential impact of using such proxies to predict the spatial dynamics of infectious diseases has received little attention despite their recurring role in informing policy and resource allocation decisions.

Therefore, better understanding both the availability of human mobility data, or lack thereof, and the impact of using mobility proxies on predicting infectious disease spread is key to prioritise future research effort. Here we focus on the African
continent which over the past decade has experienced numerous large epidemics with significant public health consequences, including multiple epidemics of Ebola [33], yellow fever [34], cholera [35], and measles [36]. We first reviewed the different data sources and mobility models that have been used to estimate human movement in Africa. We then conducted a simulation study, based on high resolution movement data from Europe, to better understand the implications of using human mobility proxies when predicting the spatial spread of infectious diseases.

**Methods**

**Search Strategy**

We carried out a scoping review of mobility data and models in Africa, adhering to the guidelines established by PRISMA extension for Scoping Reviews [37]. We searched PubMed and Web of Science on 29th August 2018 for peer reviewed literature in English on human movement data used in mobility models in Africa. We restricted our search to gravity and radiation models, which are the most commonly used in infectious disease modelling [38–42].

In short, in a gravity model the population movement between two locations is assumed to be proportional to the population sizes of the origin and distance and inversely proportional to the distance between them (Eq. (3)). The radiation model also describes how the movement between two locations may be affected by other surrounding highly populated locations (Eq. (S1)).

The search terms used were: (gravity model OR radiation model) AND human AND africa), (mobility model AND africa AND human), (gravity model AND africa), (radiation model AND africa AND human). Both abstract and full-text screening were carried out independently by two reviewers and all disagreements were resolved by consensus. Only primary research articles were eligible for inclusion. We excluded studies that were not set in Africa, did not have a spatial component, or did not use data or estimates of human mobility.

Data were systematically extracted from the studies to describe the source (e.g. census), spatial resolution, and temporal resolution of data on human movement. We extracted the location and the time period of the data for each study and whether the data were made available.

We classified as “empirical” human movement data obtained from mobile phone call detail records (CDR), micro-census data from Integrated Public Use Microdata Series (IPUMS), short-term migration data from DHS, and data on long-term migration (Global Bi-lateral Migration Database, hereafter GBMD), or on movement of refugees from the United Nations High Commissioner for Refugees (UNHCR). In addition to these generic data sources, some studies carried out surveys tailored to identify movement patterns relevant to the spread of specific diseases, e.g. overnight stays for malaria. These were also considered as empirical data sources. In contrast to these, a large number of studies used models of human mobility calibrated to data from nearby regions, or to the observed patterns of disease spread. Some studies specifically aimed to generate such mobility estimates for regions where no empirical data were available. Where a study did not rely on empirical measurements to char-
acterise mobility and instead used indirect sources such as disease spread or data from other countries, we refer to it as having used (or produced) “mobility proxies”. For those studies, we also extracted details of the mobility models used. For each country included in a study, we considered empirical data (or mobility proxy) for a given spatial resolution and for a specific time period as a discrete data set.

Simulation Study

To understand the implications of using human mobility proxies instead of empirical data when predicting the spatial spread and dynamics of infectious diseases, we carried out a simulation study comparing epidemic outcomes when using empirical movement data versus mobility proxies.

Commuting data

The settings for our simulation study were chosen based on locations where empirical mobility data were openly accessible and therefore were not set in Africa. We used commuting patterns derived by Tizzoni et al. from mobile phone call data in France and Portugal for 2007 and 2006 respectively [43]. The commuting patterns were obtained from a large sample of users’ CDR by identifying the frequently visited locations (based on proximity to mobile phone towers) of each person in the dataset. The authors assumed that the most visited place corresponded to the location where an individual lives, with the second most visited place being the location where an individual works [44]. This was then used to calculate the total numbers of people in the sample who commute between each of the 323 districts (corresponding to the ADM3 administrative units) in mainland France and the 278 municipalities (the ADM2 administrative units) in mainland Portugal. This approach also gave the number of people that both lived and worked in the same spatial unit.

We combined the commutes between and within administrative units to construct an origin-destination (O-D) movement matrix for each country. To generate population-level estimates of the total movement, we scaled the sampled movement flows between pairs of administrative units. The total number of individuals living in spatial unit $i$ and working in unit $j$ is calculated as:

$$N_{ij}^{obs} = \frac{n_{ij}}{\sum_j n_{ij} \text{pop}_i} \quad (1)$$

where $n_{ij}$ is the number of people commuting from $i$ to $j$ in the sample of mobile phone users, $\sum_j n_{ij}$ is the total number of people living in $i$ that were included in the mobile phone data sample and $\text{pop}_i$ is the population of location $i$. The scaling method in Eq. (1) ensures that the sum of each row in the O-D matrix equals the population living in that administrative unit ($\sum_j N_{ij}^{obs} = \text{pop}_i$).

Mobility proxies

In order to assess the impact of using mobility proxies in epidemic models, we constructed predicted movement matrices at the same spatial resolution as the observed
data in France and Portugal. Movement matrices can encode either the total number of individuals (referred to as O-D matrix) or the probability of moving between locations (rescaled O-D matrix). For the latter, matrix entry \( p_{ij} \) is the probability that an individual moves between spatial units \( i \) and \( j \). This is the product of the probability that i) they leave \( i \) for any destination \((1 - p_{i}^{stay})\); and ii) they move to \( j \) given that they have moved out of \( i \) (for which we use the notation \( p_{ij}^{move} \)). This overall probability can be written as:

\[
p_{ij} = \begin{cases} 
p_{i}^{stay} & \text{if } i = j \\
(1 - p_{i}^{stay})p_{ij}^{move} & \text{if } i \neq j 
\end{cases}
\]  

(2)

where \( p_{i}^{stay} \) is the probability that a person living in \( i \) also works in that same location.

Predictions of the relative flows \( p_{ij}^{move} \) were obtained from the gravity model, which posits that the flow of individuals from location \( i \) to location \( j \) is proportional to:

\[
p_{ij}^{move} \propto \frac{N_{i}^{\alpha} N_{j}^{\beta}}{d_{ij}^{\gamma}}
\]  

(3)

where \( N_{i} \) and \( N_{j} \) are the populations living in locations \( i \) and \( j \), and \( d_{ij} \) is the distance between the two locations [30, 45]. \( \alpha, \beta \) and \( \gamma \) are model parameters. We fitted the gravity model to observed movement data and then used the estimated parameters to predict \( p_{ij}^{move} \).

The gravity model in Eq. (3) cannot be used to estimate \( p_{i}^{stay} \). Therefore, \( p_{i}^{stay} \) values were taken directly from observed data and used to populate the diagonal elements of the \( p_{ij} \) matrix and to scale the off-diagonal elements as shown in Eq. (2).

We multiplied the rescaled O-D matrix by the size of the origin population \((pop_{i})\) to obtain the numbers of people predicted to move between pairs of spatial units (O-D matrix):

\[
N_{ij}^{pred} = p_{ij} pop_{i}
\]  

(4)

We generated different mobility proxies \( N_{ij}^{pred} \) by varying assumptions about the availability of movement data in a country and about \( p_{i}^{stay} \). In our central scenario presented here, we assume that movement data are unavailable for a given country and we therefore make predictions based on data from a nearby country (Fig. 1). For Portugal, we generate mobility proxies using data from France, i.e. \( p_{ij}^{move} \) estimated by a gravity model fitted to France and \( p_{i}^{stay} \) set to the average observed \( p_{i}^{stay} \) across all French administrative units. The same approach was used to generate mobility proxies for France using data from Portugal.

We explored additional scenarios where \( p_{i}^{stay} \) and \( p_{ij}^{move} \) were informed by data from either the correct or the neighbouring country and at various spatial resolutions (e.g. gravity model fitted to ADM2 but used to predict movement at ADM3 in the same country) (Suppl Tab. S1). To explore the robustness of our results to the choice of mobility model, we also considered a scenario using a radiation model to predict \( p_{ij}^{move} \).
Epidemic model

We use a stochastic discrete time SEIR metapopulation model to simulate epidemics in France and Portugal. The subpopulations in the metapopulation model are formed from each combination of home and work location. The size of each subpopulation $N_{ij}$ is the number of individuals who live in $i$ and work in $j$. $N_{ij}$ is set using either the observed data ($N_{ij}^{obs}$) or mobility proxy ($N_{ij}^{pred}$). This means that the home and work locations of individuals are fixed within our model, with movements occurring recurrently. People in each subpopulation fall into one of four compartments denoting their infection states: susceptible ($S$), exposed ($E$), infectious ($I$) or recovered ($R$).

Each subpopulation has two sets of interactions with other subpopulations. For half of each day (a day being defined as a 24-hour period), we assume homogeneous mixing between all individuals who reside in the same spatial unit (i.e. all people living in $i$) and for the other half of the day we assume homogeneous mixing between individuals who work in the same spatial unit (all people working in $j$). These assumptions are in line with previous studies that incorporated commuting data in models of infectious disease spread [43, 46].

Thus there are two forces of infection acting each day on every subpopulation in our model:

$$\lambda_{home}^i = \beta \frac{\sum_j I_{ij}}{\sum_j N_{ij}}$$
$$\lambda_{work}^j = \beta \frac{\sum_i I_{ij}}{\sum_i N_{ij}}$$

(5)

where $\beta$ is the per capita transmission rate. $\lambda_{home}^i$ therefore depends on the number of infectious people living in the same spatial unit $i$ ($\sum_j I_{ij}$). Similarly, $\lambda_{work}^j$ depends on the number of infectious people working in the same spatial unit $j$ ($\sum_i I_{ij}$). We make the simplifying assumption that movement is independent of infection state and that $\beta$ is the same in all settings.

The transitions between infection compartments at each time step ($\Delta t = 0.1$ day) are modelled stochastically as follows:

$$S_{ij} \rightarrow E_{ij} \sim \text{Binomial} (S_{ij}, 1 - \exp(-\lambda \Delta t))$$
$$E_{ij} \rightarrow I_{ij} \sim \text{Binomial} (E_{ij}, 1 - \exp(-\delta \Delta t))$$
$$I_{ij} \rightarrow R_{ij} \sim \text{Binomial} (I_{ij}, 1 - \exp(-\gamma \Delta t))$$

(6)

where $1/\delta$ and $1/\gamma$ are the mean latent and infectious periods respectively. $\lambda$ takes the value $\lambda_{home}^i$ in the first half of each day when people are at home, and $\lambda_{work}^j$ in the second half of each day when people are at work.

We simulated and compared epidemics in France and Portugal under different scenarios, where $N_{ij}$ was informed by different data. In the baseline scenario, $N_{ij}$ was set to observed data $N_{ij}^{obs}$, while in the alternative scenarios $N_{ij}$ was based on mobility proxies $N_{ij}^{pred}$ (Suppl Tab. S1). The epidemics were seeded in either the capital city (Paris and Lisbon) or a less central location (Brest in the northwest of France and Miranda do Douro in the northeast of Portugal). These locations were
chosen to explore differences in the dynamics of outbreaks originating in a prominent urban centre and a location on the edge of the commuting network.

For each scenario, we ran 100 simulated epidemics for 300 days, seeding 5 infectious cases of a flu-like pathogen and assuming that the rest of the population is fully susceptible (see Suppl Sec. 2 for further details). We restricted our analysis to epidemics which were seeded successfully (defined as simulations where > 10% of the total population was infected after 300 days).

For each scenario $s$, we summarised all the successfully seeded epidemic simulations with four metrics, two which characterise scenario $s$ only, and two which compare $s$ to the baseline:

- **Invasion time in $i$** ($t_{i}^{\text{first}}$), defined as the median time to the arrival of the first infectious case among those living in location $i$.

- **Peak time in $i$** ($t_{i}^{\text{peak}}$), defined as the median time to the epidemic peak in location $i$ (i.e. time when the number of infectious individuals living in $i$ is at a maximum).

- **Relative error in invasion time in $i$** ($\epsilon_{i}^{\text{first}}$). This was calculated as the difference between $t_{i}^{\text{first}}$ in scenario $s$ (i.e. using mobility proxy) and in the baseline (using the empirical mobility data), divided by the largest invasion times across all locations in scenario $s$: $\epsilon_{i}^{\text{first}} = \Delta t_{i}^{\text{first}} / \max_{j} (t_{j}^{\text{first}})$.

- **Invasion order similarity** ($p_{n}^{\text{first}}$), defined as the proportion of the first $n$ locations invaded in the baseline scenario that were also among the first $n$ locations invaded in scenario $s$. The invasion order was defined based on the median invasion time across all simulations with successfully seeded epidemics, i.e. $t_{i}^{\text{first}}$.

**Results**

**Scoping Review**

Of the 471 articles from the initial search, 129 were selected for abstract screening and 30 full-text articles for data extraction, all of which were published between 2007 and 2018. Across the 30 studies, we identified 150 empirical human mobility data sets and 168 mobility proxies.

**Empirical data**

Empirical data were available for 52 of 54 African countries (Fig. 2). For 51 countries, data on long-term migration patterns between countries were available, from GBMD (n = 42 countries) or data on refugee movement from UNHCR (n = 51). The GBMD provides data on international migrations between all countries for each decade in the period 1960–2000 [47]. The UNHCR collects data on the annual flows of refugees, asylum seekers and the number of internally displaced persons between a country/territory of origin and asylum [48].

7
Out of 54 countries, we found only 17 with subnational mobility data. These were informed by census (n = 14 countries), mobile phone records (CDR, n = 5, Côte d’Ivoire, Kenya, Namibia, Sierra Leone, and Senegal), social media records (n = 1) [49] or dedicated surveys (n = 5) [50, 51]. Censuses include surveys designed to measure changes in socio-demographic trends (such as internal migration) in a country, e.g. recording a change in address at ADM1 level over the last 1, 5, or 15 year period. These data are made available as either aggregate statistics (referred to as census or population-level census) or individual records (referred to as census microdata or individual-level census). Harmonised census microdata across different countries provided by the Integrated Public Use Microdata Series [52] were frequently used to quantify subnational mobility. Individual-level census data from specific geographic locations (rather than the entire country) were also available and informed more temporally and spatially resolved movements [53–55]. CDR provided the most spatially and temporally resolved data and were used for methodological research [56–58] as well as research into infectious diseases including Ebola [59], malaria [60, 61], schistosomiasis [62], and cholera [63]. Similarly, mobility data from social media use (i.e. geolocated tweets) were derived at a high spatial resolution (ADM3 in South Africa) [49]. Dedicated surveys collected mobility data relevant to the spread of a specific disease in a specific location [50, 51].

Mobility proxies

The empirical data described above were used to derive subnational mobility proxies for 44 of 54 countries in Africa, 37 of which had no empirical data informing subnational mobility (Fig. 2b). These proxies were often generated using mobility models fitted to census microdata from one or more countries, or CDR data from other African countries [4, 64, 65]. We also identified mobility proxies estimated from flight data from Europe and North America, and subsequently used to inform mobility between 140 cities across 43 countries in Africa (Fig. 2c) [66]. Overall, we identified subnational mobility proxies for 51 of 54 countries.

We found that typically, mobility proxies were not only generated using data from different locations but also extrapolated the mobility patterns from empirical data to earlier or later time periods, sometimes over a decade apart [4, 61, 63, 65, 67].

In addition to those relying on empirical movement data, mobility proxies were also inferred from indirect evidence such as spatio-temporal trends in disease incidence [68–70] or pathogen genomic information [71, 72] (Fig. 2c). Finally, mobility proxies were sometimes used to model human movement in different countries or at a different spatial scale, e.g. using gravity model parameters fitted to ADM1 unit data to describe movements between ADM2 units, in the same or another country [73].

Typically, mobility proxies quantified movement information as absolute flows over a specified time window (i.e. the number of people moving between a source and destination) [4, 64–66, 68, 69, 72], or relative flows (i.e. probability of moving from a source to a given destination, conditional on moving out of the source) [51, 57, 60, 63]. Some studies focusing on disease spread also characterised directly the probability of transmission of a pathogen between locations in a given time unit [71,
The focus was therefore to quantify the movement between locations and not the probability of remaining in the same place. However, this assumption was rarely stated explicitly, e.g. clarifying that relative flows are conditional on moving in the first place. In fact, we identified only one study that estimated the probability of not moving out of a spatial unit over a year (i.e. $p_{\text{stay}}$), albeit at a gross temporal scale [62].

In studies that used empirical data sources to calibrate mobility models, the underlying data were rarely shared with the publication, even when commercial restrictions did not prevent data sharing (55/150 empirical data sets readily available). Mobility proxies were more often available (125/168) [64–66, 73].

### Epidemic Simulation Study

We illustrate the potential implications arising from the use of mobility proxies in epidemic models. We present results for our main scenario in which mobility proxies for France are generated from a gravity model calibrated to Portuguese movement data with $p_{\text{stay}}^i$ set as the average across all Portuguese administrative units (and vice versa for Portugal). We compare each summary metric (see Methods) for simulations using mobility proxies against a baseline (using observed mobility data).

#### Performance of mobility models

The mobility proxies fitted the observed movement data moderately well in both Portugal and France ($R^2 = 0.44$ and 0.47 respectively, Fig. 3A, Fig. 4A). In a scenario where the mobility proxy was based on data from the same country, the fit was marginally better ($R^2 = 0.48$ and 0.55 respectively), and only slightly decreased when using $p_{ij}^{\text{move}}$ fitted to the other country and a local $p_{i}^{\text{stay}}$ (Suppl Tab. S4). These results suggest that, in both countries, the mobility model we used (see Eqs. (2) to (4)) is only able to explain about half of the variance in the mobility data, and performance decreases by about 10% when fitted to data from another country.

#### Impact on epidemic dynamics

Invasion times ($t_{i}^{\text{first}}$) in simulations using mobility proxies were strongly correlated with those using observed movement, especially for epidemics seeded in well-connected locations ($R^2 = 0.89$ and 0.85 for Lisboa and Paris respectively, and 0.77 and 0.78 for Miranda do Douro and Brest respectively, Fig. 3B, Fig. 4B). However invasion times tended to be under-estimated when using mobility proxies, particularly in France and more so in epidemics with peripheral seeding (Fig. 3C, Fig. 4C). In the most extreme scenario (epidemics seeded in Brest), the invasion times were predicted approximately a month too early (Fig. 4B), a considerable mismatch given the short assumed generation time (the time between infection of a case and their infector) of on average 3 days.

Despite this overall trend to under-predict invasion times, locally around peripheral seeding locations, invasion times were paradoxically over-predicted when using mobility proxies. This led to poor characterisation of the early invasion dynamics, particularly with peripheral seeding (Fig. 3E, Fig. 4E). For epidemics seeded...
in Brest and Miranda do Douro, only 60% and 55% of the first 20 patches invaded were correctly identified when using mobility proxies, but this increased to 85% and 90% for epidemics seeded in Paris and Lisboa.

The ability to predict epidemic peak times in each spatial unit ($t_{\text{peak}}^i$) when using mobility proxies was very poor in France ($R^2 = 0.05$ and 0.01 for Paris and Brest respectively, Fig. 4D) but better in Portugal ($R^2 = 0.29$ and 0.60 for Miranda do Douro and Lisboa respectively, Fig. 3D). The baseline epidemics seeded in Portugal and in Paris had limited variability in the peak times across administrative units (2-3 weeks). Using mobility proxies reproduced this small amount of variation for Portugal, but not for Paris where a synchronous peak was predicted (1 week). Similarly, although the baseline epidemic seeded in Brest had substantial variability in epidemic peaks (up to 1.5 months apart), again a synchronous peak (1 week) was predicted when using mobility proxies. The sizes of the epidemic peaks were similar across the different movement scenarios.

In both countries, and regardless of the data used to fit the mobility model, the mobility proxies had similar performance in predicting movement data. However, the resulting impact on epidemic dynamics was very different between the two countries. In Portugal, again, epidemic metrics did not differ substantially depending on the data used to fit the mobility model (Suppl Tab. S4, Figs. S5 to S7). In France, using mobility proxies calibrated to Portugal data led to underestimated invasion times, poor invasion order prediction (especially in Brest), and false synchrony in the peaks. However, when using $p_{\text{stay}}^i$ from France, all metrics substantially improved, irrespective of the country used to inform $p_{ij}^{\text{move}}$ (Suppl Tab. S4, Figs. S1 to S3).

Using the radiation model instead of the gravity model in the main scenario had a mixed impact on epidemic dynamics (Figs. S4 and S8). Overall, the performance was slightly better in France and worse in Portugal. The radiation model improved some metrics, such as the ability to predict early peak timings of the epidemics seeded in Brest, although the later peaks were predicted less well. Use of the radiation model introduced substantial heterogeneity to the Portuguese simulations that was not seen in either the baseline scenario or the additional scenarios using gravity models.

Our results demonstrate that using mobility proxies can have a substantial impact on the predicted epidemic dynamics, with complex and non-intuitive biases, which cannot be predicted when simply comparing the mobility proxies to observed movement data.

**Discussion**

Accounting for over 14% of the world’s population [74] and approximately 2.4% of the global airline passenger volume [75], Africa is a large source and sink of national and international movement of populations, and is estimated to bear half of the global burden of infectious diseases [76]. Our review revealed major gaps in the availability of empirical measurements of human mobility in Africa. This leads to mobility proxies being used in place of data, which can impair our understanding of infectious disease dynamics in Africa, as shown by our simulation study.

Although we identified empirical mobility data in all 54 African countries (Fig. 2), data informing subnational movement was only available for 17 countries. These
were mostly census data and focused on long term movement (1-5 years), which would only be relevant to the spread of pathogens with very slow progression (e.g. HIV). Despite evidence that these long term movements correlate with short-term mobility [58], census data are either not available at all or infrequently updated in many African countries (Fig. 2).

Highly temporally and spatially resolved mobility data sources [77–79], including but not limited to mobile phone records (CDR data), are increasingly used to characterise human movement and epidemic spread, most frequently in high income countries [21, 43, 80–83]. However, such data is sparse for Africa: we only found 5 African countries with available CDR data, which was subsequently used to derive mobility proxies and examine epidemic spread in 17 other African countries, and which were potentially vastly different in size, population, topography and other factors expected to influence movement patterns.

Given the lack of appropriately resolved mobility data, such use of mobility proxies, typically based on data from other countries, is very common in the African context: we identified subnational mobility proxies for 53/54 African countries. Those were sometimes based on flight data from Europe and America, highlighting the scarcity of reliable local movement data. It is worth noting that mobility proxies often focus on describing absolute or relative flows between distinct locations, but rarely attempt to quantify $p_i^{\text{stay}}$, i.e. the number or proportion of people who do not move. This is a critical ingredient to model epidemic spread, where one needs to characterise movement but also the lack thereof.

Our work shows that mobility proxies are imperfect descriptions of empirical mobility, explaining only about half of the variability in the two mobility datasets we considered. Our simulation study demonstrates that this can have substantial and non-intuitive implications on our ability to predict epidemic spread. We focused on the early invasion dynamics and the local peak dynamics, as these metrics would be critical for informing policy, through timely, appropriately scaled and optimally targeted allocation of resources and implementation of control measures [84–86].

In most scenarios explored, predicted invasion times using mobility proxies tended to be earlier than when using empirical mobility data. While this may seem less problematic than the reverse, it could lead to dismissing preventive interventions, wrongly perceived to take too long to implement given the predicted speed of invasion (e.g. building a new healthcare facility). Although the early invasion order was relatively well characterised for epidemics seeded in the capital cities, this was not the case when the epidemics were seeded in peripheral locations, and could lead to inappropriate spatial targeting of control measures. Heterogeneity in local epidemic peak times was overall well quantified in Portugal but not at all for France. Better characterising the peak heterogeneity may help better coordination of resources among different regions, e.g. allowing movement of medical staff, patients or material depending on the predicted local times of maximal epidemic burden.

These issues were overall more apparent in France than Portugal, and were much improved when using $p_i^{\text{stay}}$ from the correct country. Although $p_i^{\text{stay}}$ has received little attention by researchers, our work emphasises the importance of collecting data to estimate this parameter in all countries and at a fine spatial resolution, which can be later aggregated if needed. Indeed $p_i^{\text{stay}}$ will evidently depend on the
size of the spatial unit under consideration. For instance, ADM2 units in Portugal are on average smaller than ADM3 units in France, leading to the aforementioned underestimation of invasion times. Such data collection effort, to characterise non-movement, may be even more critical than collecting data on the destinations of the movements; indeed results from our epidemic simulations were much more sensitive to changes in $p_{i}^{\text{stay}}$ than in $p_{ij}^{\text{move}}$.

In our study, we found that using French data to inform Portugal mobility fared much better than the opposite. Population mobility in a country potentially depends on a large number of factors such as the geography of the country, population density, demographics, and distribution of economic opportunities [87–89]. Understanding the key determinants may allow us to predict to what extent and in which contexts $p_{i}^{\text{stay}}$ and $p_{ij}^{\text{move}}$ may be informed by data from other countries. However, this would only be possible if highly resolved mobility data from many countries were available, which would reduce the need for mobility proxies. Although in our simulation study, we generated mobility proxies using data from one other country, we note that some popular published proxies [64, 65] use data from multiple countries. It is unclear what the implications for disease modelling are, and whether such data pooling across countries is more appropriate than using data from a single country.

Our work has some limitations. First, since our search strategy was limited to identifying studies that used movement data in mobility models, we may have overlooked studies that primarily described mobility data and did not use any models. However, we believe it is unlikely that we have missed important data sets, as those would have been used by modelling studies captured in our search.

Second, we carried out the search in 2018. Since then, novel sources of mobility data have emerged specifically in the context of the COVID-19 pandemic. This includes community mobility reports from Google and Apple [90, 91]; however those only measure overall mobility trends which cannot directly be used to inform O-D matrices. Furthermore, many such data are being released specifically to support the COVID-19 response, and cannot replace much needed highly-resolved representative and regularly updated mobility data.

Third, we used gravity and radiation mobility models, and a multipatch compartmental model stratified by home and work location, guided by the available mobility data. More sophisticated models of mobility [92] or disease propagation [46, 93–95] could display more complex dynamics. However our simple and parsimonious approach was sufficient to highlight potential biases in epidemic spread predictions stemming from using mobility proxies instead of empirical data; this would only be magnified by using more sophisticated models.

Finally, one aspect we did not consider, which is rarely acknowledged, is that mobility proxies are outputs of statistical models, and hence carry inherent uncertainty. Such uncertainty is rarely quantified, or reported, and almost never propagated in subsequent analysis, e.g. epidemic models.

Overall, our work underscores the need for regularly updated empirical measurements of population movement within and between countries. Despite potential biases in mobile phone usage due to limited and unevenly distributed smartphone ownership [96], such data sources remain the most promising to characterise short-term human movement at high resolution. Mobile phone operators could consider...
periodic release of aggregate data sets to support public health efforts. Even though
we did not restrict our search to studies related to infectious diseases, most of the
included publications focused on their spread, highlighting the centrality of human
movement in infectious disease epidemiology. A few data sources underpinned wide-
ranging research on multiple diseases such as cholera, ebola, malaria, and schisto-
somiasis. This suggests that availability of empirical mobility data could pay large
dividends in pandemic preparedness as well as an improved understanding of spread
of diseases that are endemic or lead to recurring epidemics [97].
Figures
Figure 1: **Illustration of how movement patterns observed in France and Portugal can be used to generate mobility proxies.** The ovals on the left represent the movement patterns within each country, with the smaller dashed ovals representing different spatial units within the country. Movement between locations is shown by the blue arrows in France and purple arrows in Portugal. Darker colours indicate more movement. People that move within their own spatial unit are shown with red arrows in France and green arrows in Portugal. These movement patterns can be summarised in a movement matrix (i.e. O-D matrices, middle column) that provide information on movements between each origin-destination pair (including where the destination is the same as the origin, i.e. $p_i^{\text{stay}}$, see methods). The matrices in the right column illustrate how movement patterns in France are predicted in our central scenario based on a combination of a gravity model fitted to Portuguese movement data along with data on the probability that an individual leaves the location they live in ($1 - p_i^{\text{stay}}$). Similarly, we predict Portuguese movement using a gravity model fitted to French movement data and the average $p_i^{\text{stay}}$ observed in France.
Figure 2: **Human population mobility data and proxies in Africa.** We reviewed 30 peer-reviewed articles using human mobility data or proxies in any African country. For each country included in a study, we considered empirical data (or mobility proxy) for a given spatial resolution and for a specific time period as a discrete data set. We identified 150 empirical data sets and 168 mobility proxies. (Figure caption continued on next page)
Figure 2: (a) Locations for which empirical data on human mobility were available. Data from Global Bi-Lateral Migration Database on international long-term migration and from the United Nations High Commissioner for Refugees on the movement of refugees between countries (red circles) were used as sources of international mobility within Africa. Census data (including population-level census, census microdata, and health and demographic surveillance surveys) from 14 countries (green circles) were used as empirical sources of subnational mobility at varying spatial resolutions across countries. Call details records (CDR) were used to quantify subnational mobility patterns at a fine temporal and spatial resolution and were available for five countries (purple circles). Tailored interviews at specific sites (black circles) were also used to quantify mobility. Social media usage at a fine spatial resolution (yellow circles) from two time periods were used to characterise mobility patterns in one country. (b) Locations for which human mobility proxies derived from empirical data shown in (a) have been used in the literature. Subnational mobility proxies for 44 countries in Africa (blue circles) were generated using census, CDR, or both, using data from one or more of 14 African countries where these data were available. For 37 countries in Africa, mobility proxies were the only source of information on subnational mobility. (c) Locations for which mobility proxies were derived either from data sources not shown in (a) or indirectly from other mobility proxies. The former were obtained through mobility models calibrated to empirical data from countries outside Africa (orange circles, flight data from Europe, the United States, and Canada), or from indirect evidence of mobility such as disease spread (pink circle). The latter were characterised using mobility proxies from a different country (cyan circles). In all panels, the size of the open circles indicates the spatial scale of the data or proxy, with smaller circles indicating higher spatial resolution. The number of empirical data sets or mobility proxies derived from a given source is indicated in parenthesis in the legends.
Figure 3: Mobility predictions and epidemic simulations for Portugal. 

A) Predicted numbers moving between Portuguese ADM2 units vs observed numbers from mobile phone call records. Each purple dot represents a pair of ADM2 units. The black dots show the median predicted value (y value) versus the median observed value in each bin (where the bins are marked with grey dashed lines). Error bars show the 2.5% and 97.5% quantiles of the predicted values in that bin. B) and D) show respectively the median invasion time \( t_{first} \) and the median peak time \( t_{peak} \) amongst residents of each spatial unit (represented by a dot) across all simulations where an epidemic was successfully seeded. Plots compare the times when using observed mobility patterns (x-axis) vs the mobility proxy (y-axis) from our central scenario in the epidemic model (see Methods). The colour represents the normalised distance of each patch from the seed location (shown by the title of each grid), calculated by dividing the distance from the seed by the maximum distance from that seed. The red dashed line is where \( y = x \). 

C) Maps of the relative error in invasion time in each spatial unit when using the mobility proxy vs observed data. Blue shading indicates the invasion time occurred earlier when using the mobility proxy, while red shading indicates a later invasion time when using mobility proxies. We use a relative scale, calculated as the difference between invasion time using mobility proxy and in the baseline (using the empirical mobility data), divided by the largest invasion times across all locations when using the proxy. The seeding location is marked with a red triangle. 

E) Invasion order similarity \( p_{first} \). The y-axis shows the proportion of the first \( n \) patches invaded in an epidemic model using observed mobility that were also among the first \( n \) locations invaded when a mobility proxy was used in the epidemic model. \( n \) is the proportion shown on the x-axis. The two lines show different seed locations. The dashed red line shows the expected value if locations are chosen randomly.
Figure 4: Mobility predictions and epidemic simulations for France. Figure details as in Fig. 4 except here A) shows predicted numbers moving between French ADM3 units vs observed numbers from mobile phone call records. Each blue dot represents a pair of ADM3 units.
References


[31] F. Simini et al. “A universal model for mobility and migration patterns”. In: Nature 484.7392 (2012), pp. 96–100. ISSN: 1476-4687. DOI: 10.1038/nature10856. URL: https://doi.org/10.1038/nature10856.


[34] Yellow Fever: Recent Outbreaks in Africa. URL: https://africacdc.org/disease/yellow-fever/ (visited on 02/28/2022).


Acknowledgements

We thank Michele Tizzoni and colleagues for publicly sharing data on commuter patterns in France and Portugal, as well as for providing us with additional data on the numbers of people that do not commute. This study is partially funded by the National Institute for Health Research (NIHR) Health Protection Research Unit in Modelling and Health Economics, a partnership between UK Health Security Agency (UKHSA) and Imperial College London, in collaboration with LSHTM (grant code NIHR200908); and acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/R015600/1), which is jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat agreement and is also part of the EDCTP2 programme supported by the European Union; and acknowledges funding by Community Jameel. JW acknowledges research funding from the Wellcome Trust (grant 102169/Z/13/Z). SB acknowledges funding from the Wellcome


[90] COVID-19 Community Mobility Reports. URL: https://www.google.com/covid19/mobility/ (visited on 02/03/2022).


Supplementary information

We share two supplementary files:

1. Supplementary information on the methods and results of the simulation study
2. Data extracted from studies included in the scoping review