Call detail record aggregation methodology impacts infectious disease models informed by human mobility

Hamish Gibbs[1], Anwar Musah[1], Omar Seidu[2], William Ampofo[3], Franklin Asiedu-Bekoe[4], Jonathan Gray[5], Wole A. Adewole[5], James Cheshire[1], Michael Marks[6,7,8], Rosalind M. Eggo[9]

[1]Department of Geography, University College London, London, United Kingdom.

[2]Ghana Statistical Service, Accra, Ghana.

[3]Noguchi Memorial Institute for Medical Research, University of Ghana, Legon, Accra, Ghana.

[4]Ghana Health Service, Ministry of Health, Accra, Ghana.

[5]Flowminder Foundation, Stockholm, Sweden.

[6]Department of Clinical Research, London School of Hygiene & Tropical Medicine, London, United Kingdom.

[7]Hospital for Tropical Diseases, University College London Hospital, London, United Kingdom.

[8]Division of Infection and Immunity, University College London, London, United Kingdom.

[9]Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, United Kingdom.

Correspondence to Hamish Gibbs; Hamish.Gibbs.21@ucl.ac.uk.

Abstract

This paper demonstrates how two different methods used to calculate population-level mobility from Call Detail Records (CDR) produce varying predictions of the spread of epidemics informed by these data. Our findings are based on one CDR dataset describing inter-district movement in Ghana in 2021, produced using two different aggregation methodologies. One methodology, “all pairs,” is designed to retain long distance network connections while the other, “sequential” methodology is designed to accurately reflect the volume of travel between locations. We show how the choice of methodology feeds through models of human mobility to the predictions of a metapopulation SEIR model of disease transmission. We also show that this impact varies depending on the location of pathogen introduction and transmissibility. For central locations or highly transmissible diseases, we do not observe significant differences between aggregation methodologies on the predicted spread of disease. For less transmissible diseases or those introduced into remote locations, we find that the choice of aggregation methodology influences the speed of spatial spread as well as the size of the peak number of infections in individual districts. Our findings can help researchers and users of epidemiological models to understand how methodological choices at the level of model inputs may influence the results of models of infectious disease transmission, as well as the circumstances in which these choices do not alter model predictions.

Author Summary

Predicting the sub-national spread of infectious disease requires accurate measurements of inter-regional travel networks. Often, this information is derived from the patterns of mobile device connections to the cellular network. This travel data is then used as an input to epidemiological models of infection transmission, defining the likelihood that disease is “exported” between regions. In this paper, we use one mobile device dataset collected in Ghana in 2021, aggregated according to two different methodologies which represent different aspects of inter-regional travel. We show how the choice of aggregation methodology leads to different predicted epidemics, and highlight the conditions under which models of infection transmission may be influenced by methodological choices in the aggregation of travel data used to parameterize these models. For example, we show how aggregation methodology changes predicted epidemics for less-transmissible infections and under certain models of human

movement. We also highlight areas of relative stability, where aggregation choices do not alter predicted epidemics, such as cases where an infection is highly transmissible or is introduced into a central location.

Introduction

The volume of travel between geographic locations is widely used as an input to epidemiological models of disease transmission. Mobility data provides an approximate representation of the travel of a population by recording the movement of a sample of individuals. This sample is typically drawn from the users of a specific mobile network or application. Call Detail Records (CDR) record mobile device connections to the cellular network and are a common type of mobility data used in this context.

Metapopulation models (1) informed by CDR mobility data have been widely used to study the dynamics of infectious diseases including influenza (2), rubella (3), malaria (4,5) cholera (6) dengue fever (7) Ebola virus disease (8,9), HIV (10) and COVID-19 (11,12). Transmission models informed by CDR mobility data are particularly useful in low and middle income countries where there has been a widespread adoption of mobile devices. These data can address a lack of prior knowledge about inter-regional patterns of travel and in turn, can build greater capacity for disease surveillance and prediction. Understanding what factors influence estimates of population mobility will allow for more accurate interpretation of the results of infectious disease models which rely on human mobility data. In this paper, we focus on factors introduced at the CDR aggregation stage, where individual records from mobile subscribers are aggregated to describe population-level mobility.

CDR data used in infectious disease research is typically produced as an aggregated, censored network describing the volume of travel between pairs of locations in a specified time period. The aggregation of CDR data transforms sensitive individual-level data into a description of population-level mobility, thereby reducing the risk of disclosing personally identifiable information. Aggregation and censoring also reduces the size of a CDR dataset, making it practical for use in models of population-level infectious disease transmission.

Censoring of aggregated CDR data, where pairs of locations exchanging only a small number of travellers are removed, may create a sparse network where a significant proportion of the total network connections are missing. In order to “fill in” these missing connections, mobility models are trained using external information such as the population sizes and geographic distances between locations. The parameters of these models are evaluated given observed patterns of movement from empirical data.

Modelled mobility informed by empirical CDR data is often used to parameterize metapopulation models of infectious disease transmission (2). In a metapopulation model, individual epidemics are modelled among sub-populations in discrete spatial units, with connections between sub-populations defined by a mobility network. With this construction, the volume of mobility defines the likelihood that disease will be exported between different locations as a result of travel.

Previous research has demonstrated how estimates of infectious disease can be altered by the movement model chosen to represent population mobility, although these movement models are informed by the same input parameters (13). In our research, we investigate the impact of differences at the level of inputs to movement models, as well as the influence of the choice of movement model itself. We show how movement models are sensitive to empirical inputs and how this sensitivity leads to differing predictions of infection dynamics by subsequent epidemiological models.

Methodological choices about CDR data aggregation have important implications for the reliability of infectious disease models informed by human mobility. In our research, we focus on the extent to which methodological choices used when aggregating CDR data impact estimates of population mobility (14). We show how, given identical CDR datasets, two common methodological choices during the aggregation procedure produce different representations of an empirical movement network. The first is the “all pairs” methodology which retains the long distance network connections while inflating the number of reported travellers as a consequence; while the second is the “sequential” methodology which was designed to accurately reflect the volume of travel between locations but does not include long distance connections. These methods are implemented because of their low computational complexity in transforming large individual CDR datasets into useful representations of population-level movement.

We use CDR data collected by Vodafone Ghana and processed by the Flowminder Foundation using the open source FlowKit software (15) to investigate the impact of CDR aggregation methods on estimates of human mobility and subsequently, on the results of modelled infectious disease dynamics. This data is the result of a partnership between Ghana Statistical Service,

Ghana Ministry of Health, Ghana Health Service, Vodafone Ghana, and the Flowminder Foundation (16).

Results

Differences in estimated population movement

We found that the all pairs methodology recorded an average of 2.33 million daily trips while the sequential methodology recorded 1.35 million trips (41% fewer trips) (Figure 1). Aside from a higher overall volume of travel, the all pairs network was also more connected than the sequential network, with 13,523 connections compared to 5,805, a 57% difference. The all pairs network was also more dense (a comparison of the number of observed connections and the number of possible connections) compared to the sequential network (0.18 compared to 0.08 for the sequential network) (Table 2). The higher density of the all pairs network is likely a result of the increased number of trips and the uneven distribution of cell sites in Ghana (Supplemental Figures 2, 3).

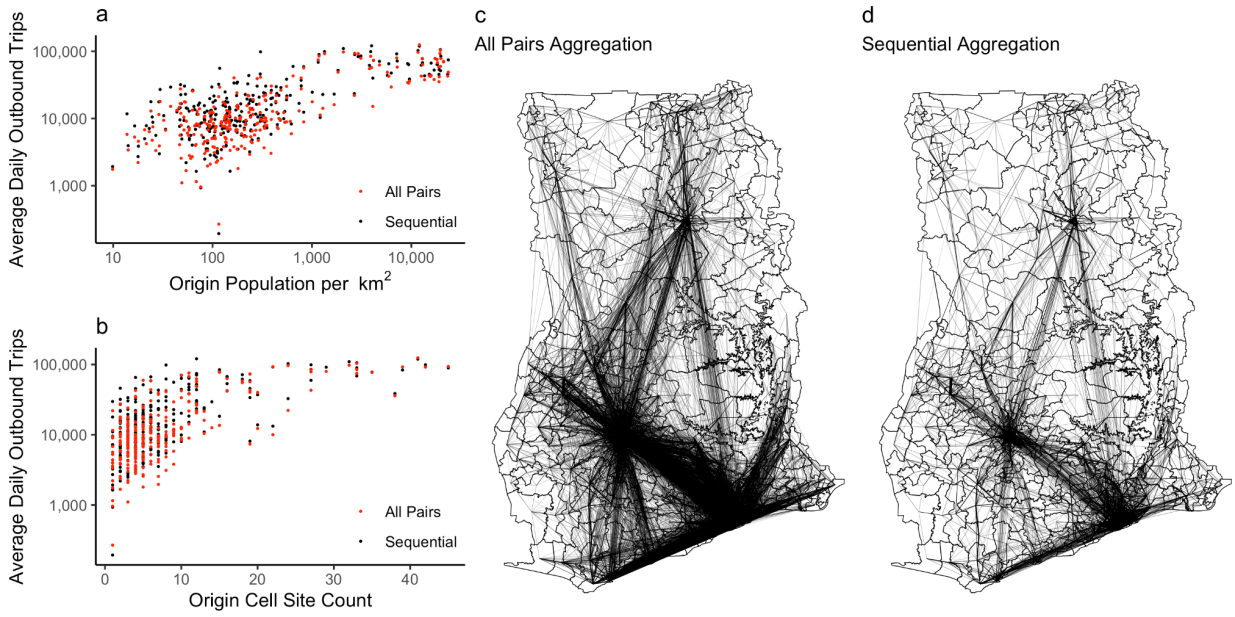

Figure 1. Aggregation methodology increases reported movement with identical underlying CDR data. *The number of recorded trips relative to a) district population and b) the number of cell sites in a district. c) The all pairs movement network and d) the sequential movement network.*

	All pairs	Sequential
Total Connections	13,523	5,805
Daily Trips	2,331,125	1,354,908
Average Degree of a district	99.8	42.8
Network Density	0.18	0.08

Table 2. Differences in observed movement caused by aggregation methodology. *Differences between two movement networks computed from the same underlying CDR data.*

Impact of aggregation on modelled human movement

Overall, each movement model reflected the differences in the empirical networks, with more connections and daily trips between districts in the all pairs methodology. However, the size of these differences varied based on the construction of the movement model (Figure 2, Table 2). The power law gravity model produced a near-fully connected network based on both aggregation methodologies but a large difference in the number of modelled trips (+46% more trips in the all pairs network). The exponential gravity model produced a less connected network overall, with greater differences in the number of connections (+39%), but somewhat smaller differences in the number of modelled trips (+43%) in the all pairs network. The radiation model produced smaller differences in the number of trips between aggregation methodologies (+38% in the all pairs network), but produced a less connected network compared to the power law gravity model or exponential gravity model (for sequential connections only).

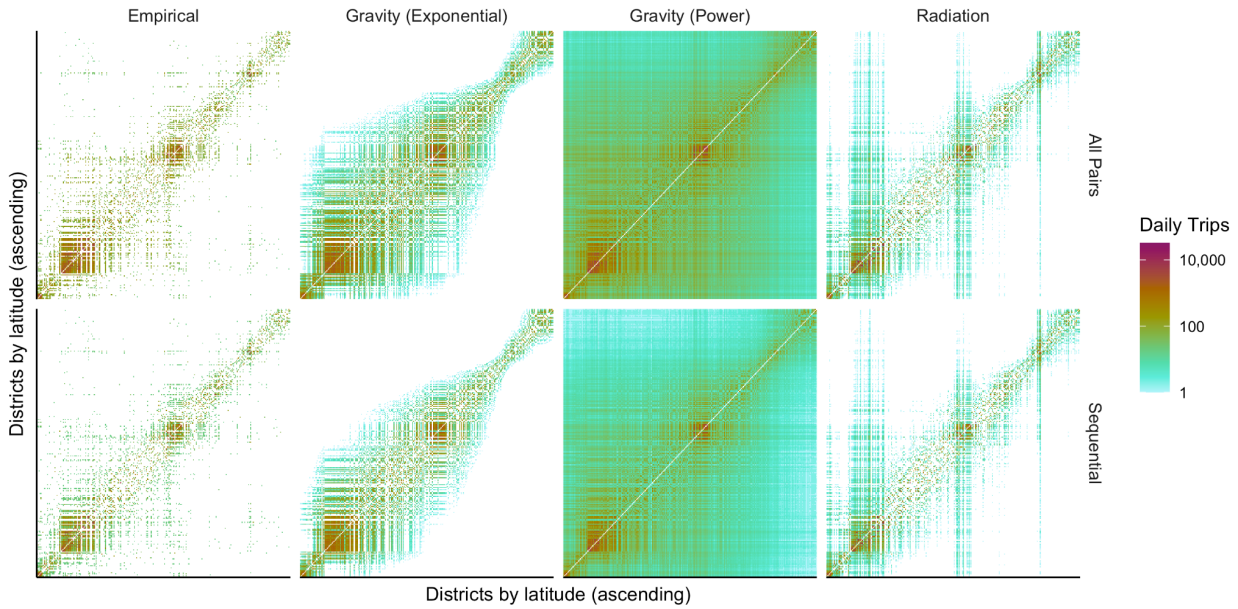

Figure 2. Empirical and modelled networks informed by different aggregation methodologies. *Comparison of the empirical movement networks (left) and modelled networks for three types of movement models. Higher numbers of travellers in the all pairs network translates into a higher number of modelled travellers for all models.*

	All pairs aggregation		**Sequential aggregation**	
Model	**Connections**	**Daily Trips**	**Connections**	**Daily Trips**
Gravity (Exponential)	22,764	2,639,262	13,979	1,503,146
Gravity (Power)	73,170	3,958,835	73,168	2,137,533
Radiation (Basic)	16,886	2,148,336	14,184	1,332,434

Table 3. Differences in travel network characteristics by movement model and aggregation methodology. *The difference in the number of modelled connections and daily trips using different models of human movement. The difference between empirical networks is reflected in predictions from each movement model.*

We compared the modelled movement networks to the underlying empirical networks, finding that the radiation model had the lowest overall Mean Absolute Percentage Error (MAPE) and highest R^2 values compared to other models, indicating closer fit with the empirical data, but produced higher Root Mean Squared Error (RMSE) compared to the other models (Table 4). This likely indicates that the radiation model had better overall fit but introduced large errors for connections between certain locations. Because models were trained on different underlying

empirical networks, these measures of model performance cannot be compared between different aggregation methodologies.

	All pairs aggregation			Sequential aggregation		
Model	**RMSE**	**MAPE**	**R^2**	**RMSE**	**MAPE**	**R^2**
Gravity (Exponential)	764.44	1.72%	0.33	659.74	2.33%	0.24
Gravity (Power)	742.73	1.21%	0.51	655.94	1.80%	0.41
Radiation (Basic)	1,292.40	0.95%	0.65	725.92	0.91%	0.68

Table 4. Evaluation of movement models for different aggregation methodologies. *The Root Mean Squared Error (RMSE), Mean Average Percentage Error (MAPE), and* R^2 *comparing modelled movement to the empirical movement networks. Note that because models were informed by different empirical networks created from different aggregation methodologies, the evaluation cannot be compared between methodologies.*

We compared the empirical networks and modelled networks, and calculated differences between aggregation methodologies for both the empirical and modelled networks predicted by each movement model (Figure 3, Supplemental Figures 4, 5, 6). Overall, we observe greater difference in the number of travellers recorded by different aggregation methodologies with respect to distance, as the length of connections increases, there is greater difference between the empirical networks. For the predictions of movement models, the difference between aggregation methodologies reflects the underlying construction of each model. This is especially evident in both gravity models (Figures 3b and 3c, Supplemental Figures 4d and 5d), whereas the difference in the radiation model (Figure 3d, Supplemental Figure 6d) more closely approximates the difference observed in the empirical networks.

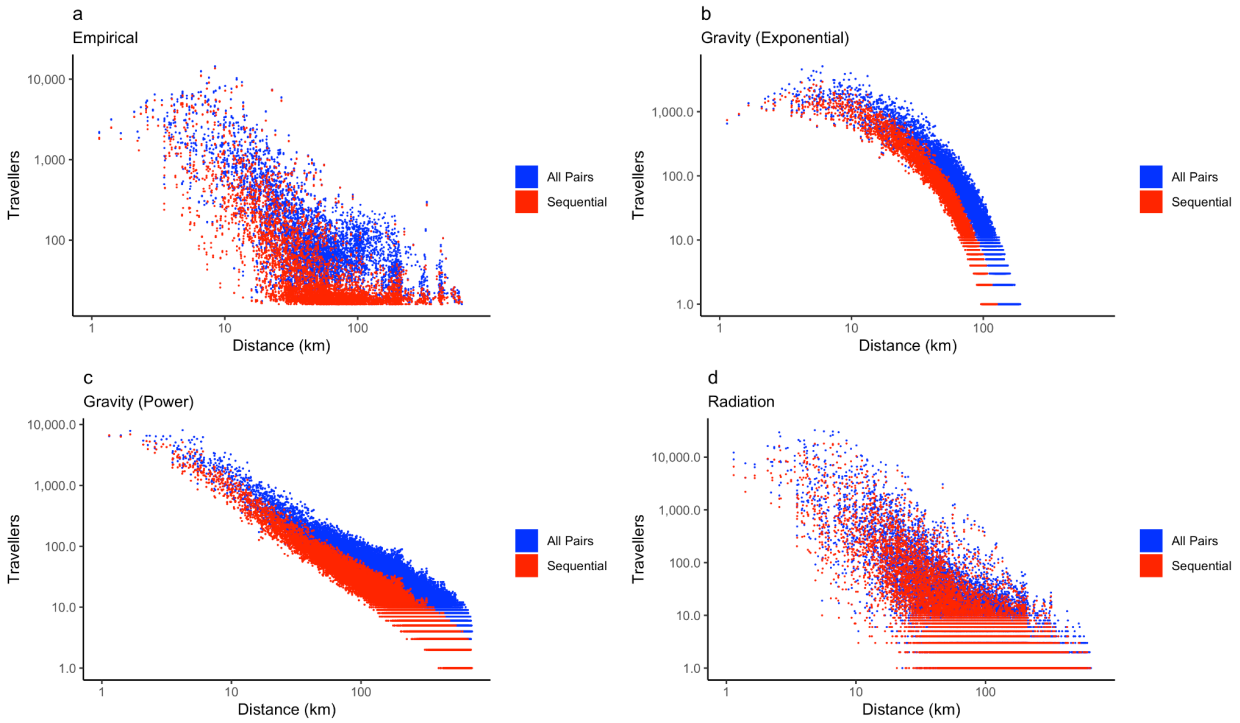

Figure 3. Comparison of empirical and modelled travel networks by aggregation methodology. *The choice of aggregation methodology results in lower number of travellers in the sequential network in a) the empirical network, b) the exponential gravity model, c) the power law gravity model, and d) the radiation model, thereby leading to an underestimation of the number of travellers compared to the all pairs network.*

Results of aggregation methodology on an SEIR metapopulation model

We found that differences in aggregation methodology influenced the predictions of the metapopulation model (Figure 4). In certain cases, the use of the all pairs aggregation methodology resulted in an earlier epidemic peak and a higher peak number of infections compared to the sequential network. The exponential gravity model, for example, produced earlier epidemic peaks and a lower peak number of infections based on the all pairs aggregation methodology compared to the sequential methodology (Figure 4, Supplemental Figure 7). However, this finding was not consistent for all movement models: the power law gravity model (Supplemental Figure 8) produced largely identical predicted epidemics irrespective of the aggregation methodology, while the radiation model (Supplemental Figure 9) produced a smaller difference in epidemics compared to the exponential gravity model. This difference likely indicates the differing importance of connections and volumes of travel in the modelled networks. The exponential gravity and radiation models had a large number of missing connections compared to the power law gravity model - indicating that connections between

locations are important for introducing the infection, where these introductions begin local chains of transmission.

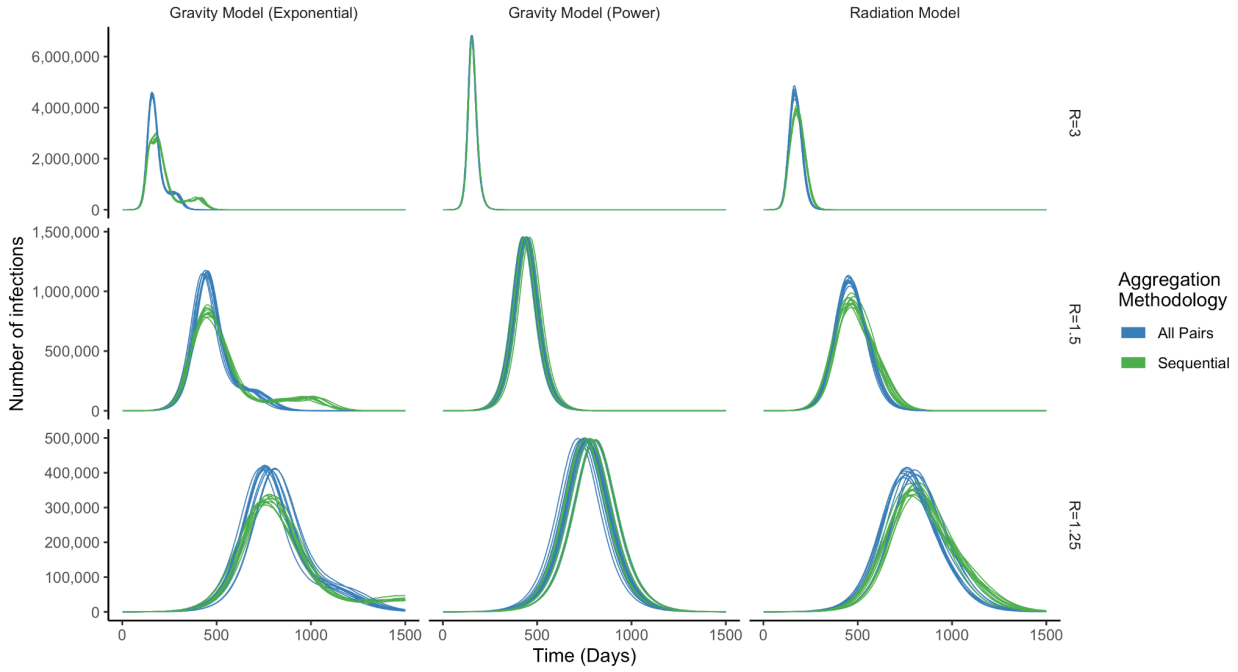

Figure 4. Comparison of modelled national epidemics by movement model. *The difference in the national epidemic modelled for different movement model constructions and different values of* R_0*. Epidemics were modelled 10 times for each combination of aggregation methodology, movement model, and* R_0*. Infection was introduced into the capital city of Accra.*

We found that the difference between aggregation methodologies was highly sensitive to the transmissibility and the location of infection introduction (Figure 5, Supplemental Figures 10, 11, 12). There was little difference between the timing and size of epidemics caused by aggregation methodology when an infection was more transmissible or was introduced into central districts in the middle and southern parts of Ghana, which include the largest cities in Ghana: Kumasi and Accra respectively. However, the choice of aggregation methodology produced a greater difference in the progression of the modelled epidemic as infections were introduced into more rural locations, particularly in the Northern parts, or were less transmissible (R_0 = 1.25, or R_0 = 1.5).

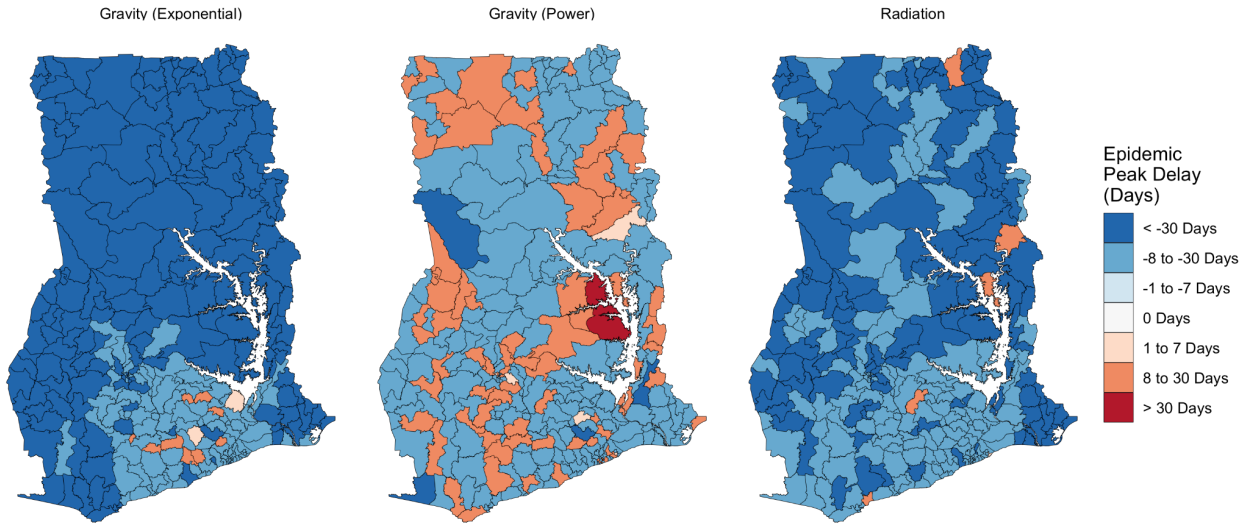

Figure 5. Difference between the peak timings of national epidemics for different movement models. *Comparison of epidemic progression between aggregation methodologies for an epidemic with* R_0 *= 1.5 seeded in each district. Negative values indicate an earlier peak in the model informed by the all pairs network. Epidemics informed by different models of human movement show how differences in aggregation methodology vary spatially as a result of aggregation methodology and because of the choice of movement model.*

We found that the choice of movement model had a notable effect on the timing of national epidemic peaks between aggregation methodologies. This reflects the differences in the underlying construction of each movement model and their subsequent impacts on disease exportation between districts. The exponential gravity model, for example, produced a more sparse movement network than the power law gravity model, which led to a spatial regularity in the timing of the epidemic peak because the effect of aggregation has been exaggerated by the modelled travel networks. By contrast, the power law gravity model, which produced a more connected network overall, produced more spatial heterogeneity in difference between epidemic peaks, particularly for less transmissible infections (R_0 = 1.25) (Supplemental Figure 11). The arrival of infection based on the sequential aggregation methodology in certain districts may reflect the greater degree of randomness in the exportation of infections in a well-connected network. The radiation model showed less spatial heterogeneity, with delayed epidemics in the Northern areas of Ghana and around Lake Volta for less transmissible infections (R_0 = 1.25) (Supplemental Figure 12).

Discussion

In this paper, we demonstrate the way that choices in the aggregation of CDR data can influence the results of models of human mobility and predictions of the spread of epidemics informed by modelled human mobility. We show how two aggregation methodologies used to respond to the COVID-19 pandemic produce different predicted epidemics in Ghana. The all pairs methodology, which produces a more densely connected network with higher volumes of travel, tends to produce an earlier epidemic with a higher peak number of infections. We show, however, that the difference between methodologies is sensitive to the transmissibility of infections, the location of infection introduction, and the choice of epidemic model. For a highly transmissible infection (R_0 = 3), or an infection introduced into an area of dense travel connections, there is little difference between aggregation methodologies because infection is spread rapidly between locations, beginning local chains of transmission which dominate the dynamics of a predicted epidemic. However, for a less transmissible infection (R_0 = 1.5 or 1.25), or for infection introduced into areas poorly connected to the travel network, infections spread more gradually, leading to a greater importance of the travel network and hence, greater importance of the aggregation methodology used to create that travel network.

A large quantity of research has used CDR data as an input to transmission models for a range of infectious diseases in different national contexts. COVID-19 provided a new challenge for the use of CDR data because of the need to provide near-real time insights about population movement in a way that used limited computational resources to produce aggregated estimates of population movement. It is in this context that the aggregation methodologies considered in this paper were developed, balancing the need to describe population movements with a need for rapid, low-complexity methods for their generation (17). This need is particularly salient in low and middle income country contexts where computational resources are often constrained and there is limited capacity for disease surveillance.

The choice of CDR aggregation methodology is typically an initial step in modelling disease transmission, and as such, there is little research concerning the impact of these apparently minor methodological choices on the predictions of epidemiological models. Previously published research has used a variety of methods to aggregate CDR data, from methods aggregating call volumes between pairs of locations, as addressed in this paper, to methods

detecting changes in home locations defined by common presence (7) or night-time location (6,18). Other research has used hybrid methods, such as recording travel to all administrative districts relative to device's home location (19). Ultimately, the choice of aggregation methodology, as well as the spatial and temporal units used for aggregation, should be chosen based on a specific hypothesis in agreement with available understanding of underlying mechanisms of disease transmission. Future research can help to identify the circumstances, such as less transmissible infection or transmission in rural areas, under which methodological choices at the stage of CDR aggregation can produce variations in the predictions of epidemic models.

Because this is a novel study, we are unable to confirm the external validity of our findings across other contexts. We therefore call for further studies that use the same methodology to generate wider analysis of disease dynamics in sub-Saharan African countries other than Ghana to assess whether the adoption of CDRs as inputs to epidemiological models will produce similar challenges as those found in the present study. However, we argue that our study is internally valid as extensive attempts were made to minimise any forms of systematic error that could potentially occur in this modelling exercise by accounting for variations in infection transmissibility, different models of human movement, and sensitivity to the location of infection introduction.

This paper focuses on the impact of methodological choices during the aggregation of CDR records. While these choices may result in important differences in observed patterns of movement, there are many other factors which influence the quantity and structure of movement captured by aggregated CDR data. These factors include uneven patterns of mobile phone usage and differences in individual travel behaviour. Some factors, like access to mobile devices or transportation may be further related to demographic characteristics like socio-economic status.

Technical factors may also alter the set of mobile devices included in CDR data. One such example is the uneven distribution of cell towers resulting in areas with minimal network connection. Mobile devices in these areas may be omitted from CDR data or may have a lower probability of generating CDRs regardless of the movement or activity of a given device. Cell towers are unevenly distributed in Ghana, clustering in population centres and along

transportation networks. Cell tower density is also correlated with population, meaning that samples of CDR data may overrepresent devices located in more populous areas.

Despite the numerous factors which influence the movement behaviour represented by CDR data, we consider that the accelerating number of travellers relative to cell tower density observed in this study points to the influence of a considerably small methodological change on the level of movement in our dataset. Other factors do not influence the difference between the empirical networks or subsequent model outputs because both aggregates were produced from the same underlying CDRs. The influence of the aggregation methodology on observed levels of movement is further supported by the association between the observed volume of movement and the theoretical prediction of movement volume.

We have shown that aggregation methodology impacts the results of movement and epidemiological models informed by CDR data and that certain aspects of these models are more sensitive to the effect of CDR aggregation. While our findings should increase researchers' caution when using CDR aggregates, this source of mobility data remains invaluable for understanding patterns of human migration, particularly in low and middle income countries like Ghana. Moreover, CDR aggregates are widely used in operational settings, as the inputs to movement and epidemiological models and to inform government decision makers. The task of human movement researchers will be to continue to improve understanding of how these data can be used to reliably describe population movements, in spite of the shortcomings of CDR data.

Materials & Methods

We used Call Detail Record (CDR) data from Vodafone Ghana, a mobile network operator which collects CDRs from subscribers to calculate billing charges. All network transactions (calls, text messages, data usage) are recorded by the CDR data, which capture the sim identifier and the identifier of the cell tower to which a device is connected. The approximate location of a device can be estimated using the known location of the connected cell towers. "Movement" of devices is derived from the sequence of locations in which a mobile device connects to a cell tower. We use CDR data aggregated to the level of districts (Administrative Level 2). There is no information on the position of a mobile device if it does not connect to the cellular network in a certain district during transit.

CDRs are aggregated to preserve the privacy of individual mobile devices and to limit the amount of storage and computational resources required to analyse CDR data. Processing CDR aggregates requires efficient algorithms for reducing billions of records to usable estimates of population movements. These algorithms rely on methodological choices that reduce the computational overhead of aggregation and allow for efficient reduction of CDR data into origin-destination (OD) matrices.

We compared one CDR dataset which was aggregated using two common aggregation methodologies: “all pairs” and “sequential” aggregation (14). Given the movement of one device between three locations A, B, and C, from A -> B, and B -> C, these methodologies produce a different number of OD pairs (Table 1).

Aggregation	**Origin-Destination Pairs**
All pairs	(A -> B) (B -> C) (A -> C)
Sequential	(A -> B) (B -> C)

Table 1. The results of different CDR aggregation methodologies. *The OD pairs produced for the movement of a single mobile device between three locations A, B, C from A -> B and B -> C. The sequential methodology produces a linear number of OD pairs while the number of pairs accelerates as the number of locations increases in the all pairs methodology.*

In Table 1 the all pairs methodology records a greater number of network connections for travel between a series of locations, which may overestimate the volume of travel in a network. However, the all pairs methodology more accurately represents long distance connections (journeys by a single individual passing through multiple locations).

Generally, using the all pairs methodology, the number of connections *e* in a network is a nonlinear function of the number of locations *n* (eq. 1).

$$e = \frac{n\,(n-1)}{2} \qquad \text{(eq. 1)}$$

With the sequential methodology, the number of connections *e* is a linear function of the number of locations *n* (eq. 2).

$$e = (n - 1) \quad \text{(eq. 2)}$$

An increase in the number of locations *n* will accelerate the number of connections *e* (and thereby increase the density of the network) in the all pairs network and will result in a constant relationship between locations *n* and connections *e* in the sequential network.

CDR Mobility Data

We used CDR mobility data collected by Vodafone Ghana and aggregated by the FlowMinder Foundation. This data was aggregated into districts (Administrative Level 2 – 271 districts). The boundaries of districts were defined by the government of Ghana. CDRs were assigned to an area based on the location of cell clusters within each region. A cell cluster is the location of a cell tower or the centroid location of a "cluster" of cell towers. In areas with a high density of cell towers, device connections may be "balanced" between multiple towers depending on network traffic and signal strength (20). This may introduce uncertainty for devices connecting to the cellular network on the boundaries of administrative areas.

We used two aggregated versions of the same underlying CDR dataset collected between February 2021 and September 2021 which recorded the daily travel between a set of origin-destination pairs (p_i, p_j) for each location *p* within a set of locations *P*. The number of travellers between locations *w* was defined as the total number of connections between pairs of locations. Pairs of locations with *w* less than 15 were removed prior to data sharing to prevent identification of individual mobile devices. The matrix of OD pairs forms a weighted directed acyclic graph of travel between locations (no information was recorded about the number of devices remaining in a given location). We calculated the average number of travellers between pairs of locations across the data collection period for use in our analysis.

Population Data

To define the population in administrative areas, we used 2020 population data from the WorldPop project (21). WorldPop population data combines population counts from national censuses with remote sensing data using Random Forest-based dasymetric redistribution to estimate the population count across a surface of 100m^2 cells. We used constrained population

estimates, meaning that population counts match population counts from the Ghana Statistical Service, but were not adjusted to match UN national population estimates. We aggregated population estimates to administrative areas in Ghana using a spatial intersection of administrative boundaries with the population surface.

Movement Models

The empirical movement matrices used in this study included missing values where travel between pairs of locations did not exceed the censoring threshold of 15 trips during the study period. To fill in these missing connections between locations, we used two common formulations of gravity models to model the movement network for connections between all locations, and the radiation model. This comparison allowed us to assess sensitivity of our findings to the choice of mobility model.

First, we used a power law gravity model defining the number of trips $\lambda_{i,j}$ between locations *i* and *j* as a function of the population size of the origin N_i, the destination N_j, and the distance between the origin and destination $d_{i,j}$ (eq. 3).

$$\lambda_{i,j} = \theta * \left(\frac{N_i^{\omega_1} N_j^{\omega_2}}{d_{i,j}^{\gamma}}\right) \quad \text{(eq. 3)}$$

In this model, travel between locations is defined by four parameters: a scaling parameter θ, and weight parameters ω_1, ω_2, and γ, which alter the contributions of origin populations, destination populations, and distance respectively.

Second, we used an exponential gravity model defining the number of trips $\lambda_{i,j}$ between locations *i* and *j* using four parameters: a scaling parameter θ, and weight parameters ω_1, ω_2, and δ, which alter the contributions of origin populations, destination populations, and distance respectively (eq. 4).

$$\lambda_{i,j} = \theta * \left(\frac{N_i^{\omega_1} N_j^{\omega_2}}{e^{\frac{d_{i,j}}{\delta}}}\right) \quad \text{(eq. 4)}$$

Finally, we used a basic radiation model defining the number of trips $\lambda_{i,j}$ between locations *i* and *j* as a function of the population size of the origin N_i, the destination N_j, the total number of trips leaving the origin M_i and the population surrounding the origin $s_{i,j}$ defined by the population within the radius $r_{i,j}$ (eq. 5).

$$\lambda_{i,j} = M_i \frac{N_i N_j}{(N_i + s_{i,j})(N_i + N_j + s_{i,j})} \quad \text{(eq. 5)}$$

We fitted all mobility models using the Mobility (22) and rjags (23) R packages. Both gravity models were fitted to the empirical movement matrices using Markov Chain Monte Carlo (MCMC) parameter estimation. Both gravity models were fitted as likelihood functions with the number of trips specified as a Poisson distribution; whereas the weak informative prior distributions for the parameters θ, ω_1, and ω_2, were defined by the Gamma distribution with shape and scale of 0.001 for parameter θ and with shape and scale of 1 for ω_1, and ω_2. The prior distribution of the parameter δ was modelled using a normal distribution truncated at 0 with mean and standard deviation calculated from the distance matrix. MCMC training was conducted using 4 chains of 50,000 samples each, with a burn-in of 10,000 samples. We assessed convergence using the $\widehat{R}$ convergence diagnostic, requiring a threshold where all parameters were deemed valid with $\widehat{R}$ less than 1.05.

We compared empirical and modelled networks by the total number of edges (connections) in the network, the total number of network trips (the sum of weights along each edge), the average node degree (the average number of edges connected to each node), and the network density (the ratio of the number of network edges compared to the number of possible edges). We also compared the performance of each model against the empirical data using Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and R^2. We assessed the quality of model predictions by identifying the model with the lowest RMSE and MAPE, and highest R^2, indicating a close fit with the empirical travel network while minimising model error.

Epidemiological Modelling

We modelled the spread of infection using a stochastic metapopulation SEIR model (24–26). The progression of the epidemic was modelled in discrete time intervals in individual subpopulations and infectious individuals were exported between subpopulations relative to the weight of the movement network connecting these populations. In the model, the probability that infections will be exported between subpopulations reflects the size of the epidemic within subpopulations as well as the volume of connections to other subpopulations. We compared the results of a model in which we vary the CDR aggregation method, and all other model parameters remain constant.

We used a stochastic SEIR model implemented in the *R* package *SimInf* (26). This model simulates an epidemic by modelling the transition of individuals in a population between compartments (Susceptible, Exposed, Infected, Removed). The model is stochastic, meaning that transitions between compartments are modelled through a random count measure and infection states for each subpopulation form a Continuous Time Markov Chain. Population movements are scheduled during model compilation, where the number of individuals travelling between subpopulations is defined by the average number of individuals travelling between pairs of locations per day across the study period. The probability of travel between subpopulations is equal across compartments.

Our model assumes constant rates of replacement (births) and mortality (deaths) within subpopulations during the study period. Although this assumption does not reflect real population characteristics, there is not sufficient data to estimate the rate of population change in Ghana during the study period. The model also assumes a uniform contact rate among members of a subpopulation. In reality, within-population contact rates vary relative to age structure and other demographic factors which are not captured by our model. The inclusion of some age-structure adjustment within population-units, as opposed to population units only, would reduce the uncertainty and fine-tune the predictions from this analysis.

To assess whether CDR aggregation methodology impacts modelled infection spread, we run the model for three values of R_0: 1.25, 1.5, 3.0. R_0 is an epidemiological parameter which defines the average number of infections arising from a single infection in a fully susceptible population. We chose these values of R_0 to simulate a COVID-19 type infection with R_0 between

1 and 3. Because R_0 is not a model input parameter, we vary the transmission rate β, given a constant recovery rate, γ.

We compare the influence of an introduction location by introducing 100 index infections into all districts in Ghana. Because our model is stochastic, meaning that the results of the model vary in part on random probabilities that individuals will transition between compartments through time, we also performed a sensitivity analysis for 5 districts, introducing index infections into 5 representative locations in Ghana's 3 major metropolitan areas: Accra (Greater-Accra Region), Kumasi (Ashanti Region), Tamale (Northern Region), and two rural districts: Lawra (Upper West Region) and Nkwanta South (Oti Region) (Supplemental Figure 1) for which we sample 10 introductions to assess the stability of the predicted spread of infection.

Data Availability

CDR mobility data used in this study was provided by Vodafone Ghana in partnership with the Flowminder Foundation and Ghana Statistical Service. This data is available to researchers by application. Use of this data was approved by the LSHTM Research Committee (Ref: 22477) and the Noguchi Memorial Institute of Medical Research (Ref: 048/20-21). Population data was publicly available 2020 constrained population counts in 100m grid cells (not UN-adjusted), downloaded from the WorldPop project. Code used in this study is available from: https://github.com/hamishgibbs/ghana_cdr_aggregation.

Acknowledgments

The following funding sources are acknowledged as providing funding for the named authors. EDCTP2 (RIA2020EF-2983-CSIGN: HPG, RME, MM). HDR UK (MR/S003975/1: RME). This research was partly funded by the National Institute for Health Research (NIHR) using UK aid from the UK Government to support global health research. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the UK Department of Health and Social Care (NIHR200908: RME). ESRC UBEL Doctoral Training Partnership (HPG). UK MRC (MC_PC_19065: RME).

Bibliography

1. Riley S. Large-scale spatial-transmission models of infectious disease. Science. 2007 Jun 1;316(5829):1298–301.
2. Tizzoni M, Bajardi P, Decuyper A, King GKK, Schneider CM, Blondel V, et al. On the Use of Human Mobility Proxies for Modeling Epidemics. PLOS Comput Biol. 2014 Jul 10;10(7):e1003716.
3. Wesolowski A, Metcalf CJE, Eagle N, Kombich J, Grenfell BT, Bjørnstad ON, et al. Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. Proc Natl Acad Sci. 2015 Sep;112(35):11114–9.
4. Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, et al. Quantifying the Impact of Human Mobility on Malaria. Science. 2012 Oct 12;338(6104):267–70.
5. Marshall JM, Wu SL, Sanchez C. HM, Kiware SS, Ndhlovu M, Ouédraogo AL, et al. Mathematical models of human mobility of relevance to malaria transmission in Africa. Sci Rep. 2018 May 16;8(1):7713.
6. Bengtsson L, Gaudart J, Lu X, Moore S, Wetter E, Sallah K, et al. Using Mobile Phone Data to Predict the Spatial Spread of Cholera. Sci Rep. 2015 Mar 9;5(1):8923.
7. Wesolowski A, Qureshi T, Boni MF, Sundsøy PR, Johansson MA, Rasheed SB, et al. Impact of human mobility on the emergence of dengue epidemics in Pakistan. Proc Natl Acad Sci. 2015 Sep 22;112(38):11887–92.
8. Wesolowski A, Buckee CO, Bengtsson L, Wetter E, Lu X, Tatem AJ. Commentary: Containing the Ebola Outbreak - the Potential and Challenge of Mobile Network Data. PLoS Curr. 2014 Sep 29;6:ecurrents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e.
9. Halloran ME, Vespignani A, Bharti N, Feldstein LR, Alexander KA, Ferrari M, et al. Ebola: Mobility data. Science. 2014 Oct 24;346(6208):433–433.
10. Brdar S, Gavrić K, Ćulibrk D, Crnojević V. Unveiling Spatial Epidemiology of HIV with Mobile Phone Data. Sci Rep. 2016 Jan 13;6(1):19342.
11. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science. 2020 Apr 24;368(6489):395–400.
12. Chang S, Pierson E, Koh PW, Gerardin J, Redbird B, Grusky D, et al. Mobility network models of COVID-19 explain inequities and inform reopening. Nature. 2021 Jan;589(7840):82–7.
13. Daniel T. Citron, Carlos A. Guerra, Carlos A. Guerra, Andrew J. Dolgert, Sean L. Wu, John M. Henry, et al. Comparing metapopulation dynamics of infectious diseases under different models of human movement. Proc Natl Acad Sci U S A. 2021 May 4;118(18).
14. Flowminder COVID-19 Resources - Mobility indicators: Descriptions [Internet]. 2020 [cited 2022 Mar 29]. Available from: https://covid19.flowminder.org/mobility-indicators/mobility-indicators-descriptions
15. FlowKit. FlowKit [Internet]. 2022 [cited 2022 Nov 3]. Available from: https://flowkit.xyz/master/
16. Li T, Bowers R, Seidu O, Akoto-Bamfo G, Bessah D, Owusu V, et al. Analysis of call detail records to inform the COVID-19 response in Ghana—opportunities and challenges. Data Policy. 2021 ed;3:e11.
17. Flowminder. COVID-19 Resources - Our approach [Internet]. [cited 2022 Oct 24]. Available from: https://covid19.flowminder.org/our-approach
18. de Monasterio J, Salles A, Lang C, Weinberg D, Minnoni M, Travizano M, et al. Analyzing the spread of chagas disease with mobile phone data. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). 2016. p. 607–12.
19. Green D, Moszczynski M, Asbah S, Morgan C, Klyn B, Foutry G, et al. Using mobile phone

data for epidemic response in low resource settings—A case study of COVID-19 in Malawi. Data Policy. 2021 ed;3:e19.
20. Sharma A, Roy A, Ghosal S, Chaki R, Bhattacharya U. Load balancing in Cellular Network: A review. In: 2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12). 2012. p. 1–5.
21. WorldPop. The spatial distribution of population in 2020, Ghana. Glob High Resolut Popul Denominat Proj.
22. Tools for Analyzing Human Mobility Data [Internet]. [cited 2022 Aug 8]. Available from: https://covid-19-mobility-data-network.github.io/mobility/
23. Plummer M, Stukalov A, Denwood M. rjags: Bayesian Graphical Models using MCMC [Internet]. 2022 [cited 2022 Aug 8]. Available from: https://CRAN.R-project.org/package=rjags
24. Fast event-based epidemiological simulations on national scales - Pavol Bauer, Stefan Engblom, Stefan Widgren, 2016 [Internet]. [cited 2022 May 17]. Available from: https://journals.sagepub.com/doi/10.1177/1094342016635723
25. Engblom S, Widgren S. Chapter 11 - Data-Driven Computational Disease Spread Modeling: From Measurement to Parametrization and Control. In: Srinivasa Rao ASR, Pyne S, Rao CR, editors. Handbook of Statistics [Internet]. Elsevier; 2017 [cited 2022 May 17]. p. 305–28. (Disease Modelling and Public Health, Part A; vol. 36). Available from: https://www.sciencedirect.com/science/article/pii/S0169716117300056
26. Widgren S, Eriksson R, Engblom S, Bauer P, Rosendal T, of 'kvec.h' .) AC (Author. SimInf: A Framework for Data-Driven Stochastic Disease Spread Simulations [Internet]. 2022 [cited 2022 May 17]. Available from: https://CRAN.R-project.org/package=SimInf