Estimating household contact matrices structure from easily collectable metadata

Lorenzo Dall'Amico[1,*], Jackie Kleynhans[2,3], Laetitia Gauvin[1], Michele Tizzoni[1,4]
Mvuyo Makhasi[2], Nicole Wolter[2,5], Brigitte Language[6], Ryan G. Wagner[7]
Cheryl Cohen[2,3], Stefano Tempia[2,3], Ciro Cattuto[1,8]

[1] ISI Foundation, Turin, 10126, Italy
[2]National Institute for Communicable Diseases of the National Health Laboratory Service, Johannesburg, South Africa
[3]School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[4]Department of Sociology and Social Research, University of Trento, Trento, Italy
[5]School of Pathology, University of the Witwatersrand, Johannesburg, South Africa
[6]Unit for Environmental Science and Management, Climatology Research Group,
North-West University, Potchefstroom, South Africa
[7]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt)
[8]University of Turin, Department of Informatics, Turin, 10124, Italy
[*]Corresponding to: lorenzo.dallamico@isi.it

October 13, 2022

Abstract

The ability of developing accurate models plays a pivotal role in understanding infectious diseases dynamics and designing effective measures for epidemic mitigation. Contact matrices are of crucial importance to quantify the interaction between age groups, but their estimation is a rather time and effort consuming task. In this article we show that they can reliably be estimated combining a context-independent model with easily collectable data, such as co-presence and an age activity parameter. The proposed model is tested on high resolution proximity data collected in a rural and an urban village in South Africa. Given its simplicity and interpretability, we expect our method to be easily applied to other contexts as well and we identify relevant questions that need to be addressed during the data collection procedure.

1 Introduction

Infectious diseases such as COVID-19 and influenza are rapidly transmitted through close proximity contacts [1] and the modeling thereof is a problem of major interest for public health. The design of effective non-pharmaceutical interventions to mitigate the epidemic spreading often relies on models capable to predict the future or to reconstruct the past of the epidemic's state, see for instance [2, 3, 4, 5, 6].

Households represent the minimal unit of disease transmission and play a fundamental role in determining the evolution of a viral spread [7]. Empirical evidences suggest that, especially at the household level, the commonly adopted homogeneous mixing hypothesis is insufficient to faithfully explain contagion [8, 9, 10]. On the contrary, it is necessary to account for age-dependent *contact matrices* that represent the diversities – across different age classes – in the frequency of contacts as well as in the transmission parameters [11, 12, 13].

Determining *household contact matrices* (HCM) is technically challenging, especially in low-resource sub-Saharan African countries with high infectious diseases burden and where the data collection is still very limited [14, 15, 16, 17, 18]. The most commonly adopted strategy relies on surveys in which the participants have to self-report their contacts in terms of number, duration and (presumed) age of the

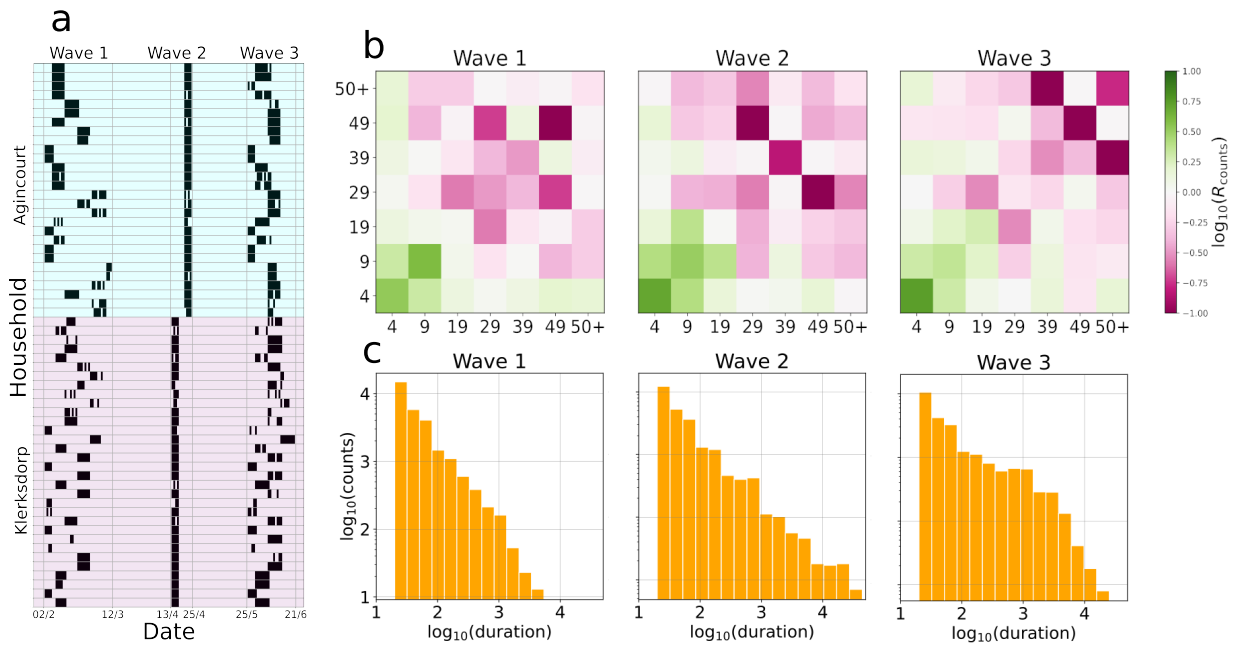

Figure 1: **Properties of the measured data**. **a**: *measurement profiles for the* 60 *selected households*. Each row corresponds to a household with a different background color distinguishing the rural site (cyan on top) from the urban site (purple on the bottom). Time is displayed on the x axis and dates are reported in the `day/month` format. Vertical gray lines correspond to the beginning and end of each deployment. A black dot indicates that at least one contact was measured for a particular (*household-deployment*) pair, while a cyan/purple one that no contact has been recorded. **b**: *normalized contact matrix across the three deployments*. The color code refers to the values of the logarithm of R_{counts} whose entries are proportional to the ratio between the number of contacts and the number of possible interacting pairs, setting the mean of R_{counts} to 1. The axis correspond to the age groups and the number reported indicates the highest age of each group. **c**: *contact duration distribution across the three deployments* in logarithmic scale.

interacting individual [19, 20, 21, 22]. Known limitations of this technique include under-reporting of contacts and overestimation of their durations [23, 24].

In this study HCM are obtained from high resolution in space and time proximity measurements, encoding the sequence of contacts among a group of selected participants. The data are collected with the proximity sensors developed by the `SocioPatterns` collaboration (http://www.sociopatterns.org/). These measurements allow one to study and model human dynamics [14, 25, 26, 27, 28, 29] and directly estimate HCM by aggregating individuals' contacts across time.

The logistic and economic effort needed to set up proximity measurements is, however, one of the main challenges associated with this data collection strategy, especially in the low-income countries where it is most needed. Consequently, identifying simple and generalizable techniques to improve it is a problem of significant importance. More specifically, in this paper we address the question *"to what extent can HCM, obtained from high-resolution measurements, be estimated from more easily collectable data?"*.

We provide an answer based on the dataset of the PHIRST study [30, 31] (details in the supplementary material), a 3-year long experiment conducted in South Africa, aimed at providing reliable data-driven guidance for public health measures to limit viral transmission [30, 32, 33, 34, 35, 36, 37, 38]. The proximity sensors were deployed three times for approximately one week in the period from February to June 2018 in Agincourt (a rural site) and Klerksdorp (an urban site), as shown in Figure 1a. The proximity contact measurements are further equipped with metadata on the individuals participating in the experiment. In the remainder we consider a total of 60 households (28 in Agincourt and 32 in Klerksdorp) with 307 individuals (151 in Agincourt and 156 in Klerksdorp) for which the measurement quality was sufficiently good in all three deployments (for further details, refer to the supplementary material).

Some of the most popular methods to estimate contact matrices rely on the demographic properties of the population under study [39, 21, 20]. In this paper we show that although these properties play a

R_{counts}	First	Second	Third
First	1	0.94	0.89
Second	0.94	1	0.94
Third	0.89	0.94	1

R_{sec}	First	Second	Third
First	1	0.85	0.87
Second	0.85	1	0.87
Third	0.87	0.87	1

Table 1: **Contact matrix similarity across the deployments**. Cosine similarity between the measured normalized contact matrices R_{counts} (left) and R_{sec} (right) in the three deployments.

crucial role in determining the HCM structure, they are insufficient to accurately model contacts and further age-dependent parameters must be deployed. We propose a model that approximates the measured HCM with a cosine similarity equal to 0.96 and 0.98 in the two sites, respectively and we claim that its optimal parametrization can be estimated from easily collectable metadata.

2 Results

With the PHIRST dataset at hand, we provide a simple model to approximate the HCM that combines three age-dependent parameters: the number of individuals per age group, the daily activity pattern and an intensity of activity factor. Below, we propose some questions to be addressed to estimate the two latter quantities. Notably, all the parameters involved in the model only depend on a single age class and not on the interactions between pairs of age classes, as it is commonly required in self-reporting surveys. This allows us to decrease the number of parameters to be estimated from order of n_{age}^2 to n_{age}, where n_{age} is the number of age classes.

We fit our model on the PHIRST dataset and show that the intensity of activity per age class significantly correlates with the presence of a main activity outside the household and with the individual's total number of interactions, averaged over all members in the same age group. This suggests that the parameters of our model can be estimated from more easily observable behaviors that quantify social interaction. We hence propose some questions to address when devising an experiment to measure contact matrices in order to quantify these parameters:

- *How much time do you typically spend at home in each hour of the day?*
- *How much of this time do you typically spend in isolation?*
- *With how many people do you engage in face-to-face interaction per day?*

These kind of questions would permit to calibrate the age dependent parameters to faithfully estimate the contact matrices, as shown in the remainder.

Let us now set a minimal notation to describe our main results. Contact matrices are square and symmetric, of size n_{age}, the number of age bins considered. Here the age groups are divided into $[0-4, 5-9, 10-19, 20-29, 30-39, 40-49, 50+]$ years: the finer grain of younger ages is justified by the rapid decay of the age distribution shown in the supplementary material. With the notation $C^{(h,d)}_{\text{counts/sec}}$ we refer to the contact matrix of household h in deployment d, storing the counts/time of interaction between pairs of age groups, while $R^{(h,d)}_{\text{counts/sec}}$ is its normalized version in which the entry (ab) is divided by the probability of random encounter between the age groups a and b. The notation $M^{(\mathcal{X})}$ (with $M = C_{\text{counts/sec}}, R_{\text{counts/sec}}$) refers to the aggregated contact matrix over a set of n *(household-deployment)* pairs $\mathcal{X} = \{(h_1, d_1), \ldots, (h_n, d_n)\}$. For simplicity of notation, whenever a superscript or a subscript is omitted it means that the sentence that has been formulated is valid for any value that the super/subscript can take.

In Section 4.1 we provide a formal definition of all contact matrices appearing in the paper. We now inspect the basic properties of these contact matrices as measured by the sensors.

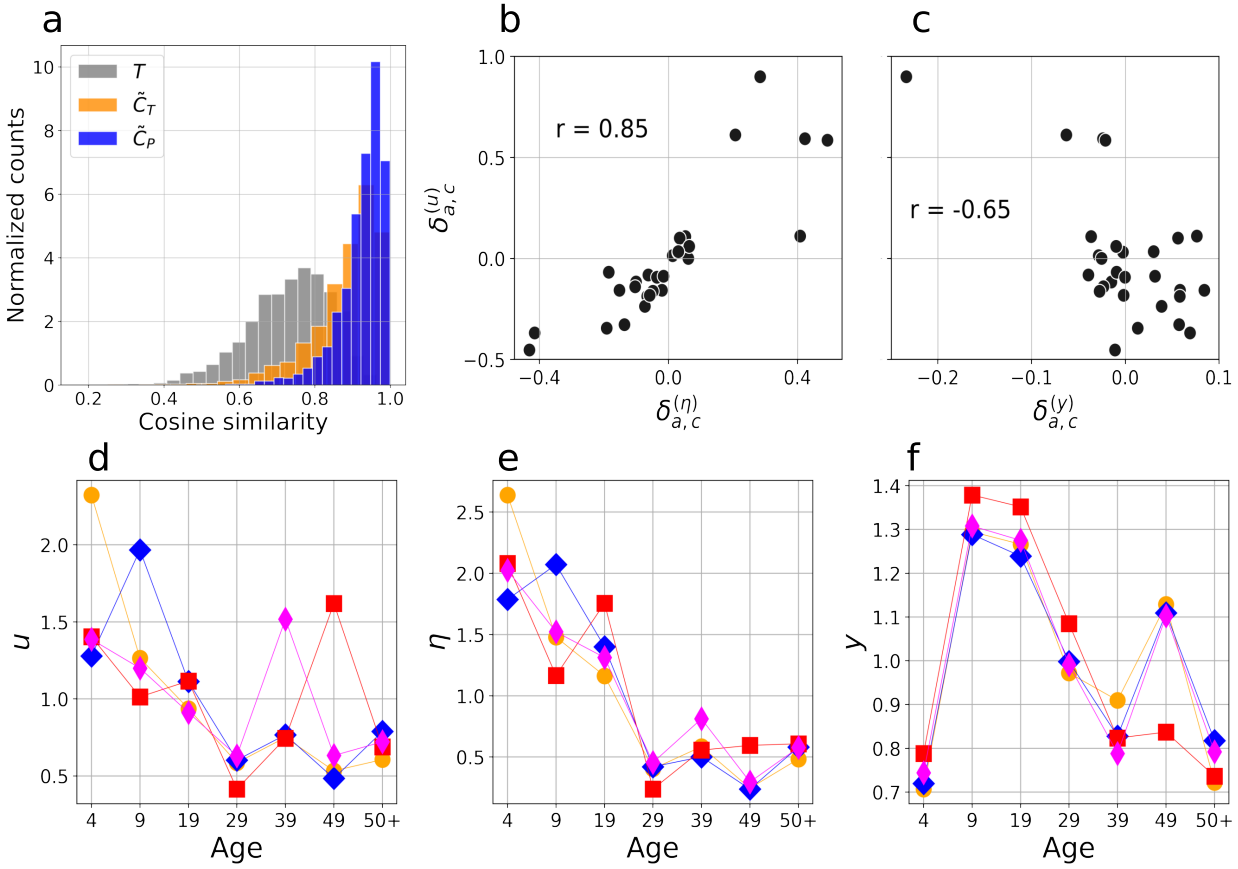

Figure 2: **Test of the model for household interaction**. **a**: *histogram of the cosine similarity between* C_{counts} *and its estimators.* The gray curve corresponds to the histogram over the 2500 realization of $\mathcal{X}$ using T as an estimator of C_{counts}. The orange curve is obtained with the first order model of Equation (1), while the blue curve corresponds to the second order model of Section (2.3). **b, c**: *correlation between the fluctuations of the activity* $\delta^{(u)}$*, the group average degree* $\delta^{(\eta)}$ *and the presence of a major occupation outside the house* $\delta^{(y)}$. The quantities $\delta_{a,c}$ are defined in Equation (2). The Pearson correlation coefficient $\mathbf{r}$ is reported in text. **d, e, f**: *averages of* $\boldsymbol{u}, \boldsymbol{\eta}, \boldsymbol{y}$ *for each group, normalized by its mean.* Each group has a color and a marker: the legend is consistent across the three plots.

2.1 Basic data properties

Since we chose to work with the same set of households in all three deployments, changes across time of the measured HCM can be due to: variations in the line-up as a consequence of migrations of some individuals or technical problems with specific sensors; the random nature of human interaction; a seasonality effect. Changes in the line-up are easily controllable, since they are accounted for in our records on deployments. They can be kept in consideration by comparing R instead of C. Table 1 precisely shows the cosine similarity between R_{counts} (left) and R_{sec} (right) for $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3$ being the set of all households in the three deployments.

The table shows high similarity values for R_{counts}, indicating that the structure of the contact matrix does not vary a lot across the three deployments. Smaller values are instead obtained by R_{sec} implying that the seasonality effect majorly involves the duration (rather than the structure) of the contacts. This evidence agrees with the distribution of the contact duration shown in Figure 1c which follows the well-known power-law decay [40], broadening in the third deployment when south-African winter is approaching. More quantitatively, we computed the 99^{th} percentile for the three distributions that is approximately 12 minutes in the first deployment, 27 in the second and 60 in the last.

Figure 1b displays instead the matrix $\log(R_{\text{counts}})$ (normalized so that its mean equals 1) across the three deployments, showing that younger age groups tend to interact more, regardless of the age group they are interacting with.

Based on these evidences, we attempt to model the matrix C_{counts} whose behavior is more predictable

than C_{sec} and rather constant across the three deployments. Given the result of Table 1, the deployments are treated as three independent, equally reliable measurements of the HCMs.

2.2 A first order model for household interaction

Let $T^{(\mathcal{X})}$ denote the "random encounter contact matrix" (formally defined in Section 4.1), whose entries indicate the probability of encounter between pairs of age groups, given the demographic data [39]. We define $\tilde{C}_T^{(\mathcal{X})}$ as

$$\left(\tilde{C}_T^{(\mathcal{X})}\right)_{ab} = T_{ab}^{(\mathcal{X})} u_a^{(\mathcal{X})} u_b^{(\mathcal{X})}, \tag{1}$$

where $\boldsymbol{u}^{(\mathcal{X})} \in \mathbb{R}^{n_{\text{age}}}$ is a parameter that represents the activity of each age group. According to this model, many interactions are expected to be observed when many members are present (large values of $T_{ab}^{(\mathcal{X})}$) and when they correspond to highly active age groups, such as $[0-4, 5-9]$, as per Figure 1b.

We deploy the following steps to test our model, as detailed and motivated in Section 4.2. We independently sample 2500 sets $\mathcal{X}$ of 8 *(household-deployment)* pairs out of the 180 available. For each sampled $\mathcal{X}$ we compute the vector $\boldsymbol{u}^{(\mathcal{X})}$ that best approximates $C_{\text{counts}}^{(\mathcal{X})}$. The entries of this vector contain the activity of each age group for the set $\mathcal{X}$. Figure 2a displays the histogram of the cosine similarity between the approximation $\tilde{C}_T^{(\mathcal{X})}$ and the measured matrix $C_{\text{counts}}^{(\mathcal{X})}$ and evidences a good agreement between the two matrices with a cosine similarity equal to 0.9 or larger for 53% of the data. Figure 2 further shows the same histogram for $T^{(\mathcal{X})}$ being used as an estimator of $C_{\text{counts}}^{(\mathcal{X})}$. This purely demographic model is much less accurate and reaches a cosine similarity greater than 0.9 for only 7% of the data.

Besides the goodness of the approximation itself, our main interest is to assess whether the vector $\boldsymbol{u}^{(\mathcal{X})}$ can be estimated from easily observable quantities. To do so, for each sampled $\mathcal{X}$ we further compute the vector $\boldsymbol{\eta}^{(\mathcal{X})} \in \mathbb{R}^{n_{\text{age}}}$, indicating the daily age group average number of interactions per individual. Intuitively, $\boldsymbol{u}^{(\mathcal{X})}$ and $\boldsymbol{\eta}^{(\mathcal{X})}$ should correlate: a higher activity should be observed when people are more active. Note that $\boldsymbol{\eta}$ aggregates *all* individual's contacts and is oblivious of the age group binning.

To investigate the correlation between $\boldsymbol{u}$ and $\boldsymbol{\eta}$, we divide the sets $\mathcal{X}$ according to their activity vector representation $\boldsymbol{u}^{(\mathcal{X})}$ into $k = 4$ groups. This appears to be the most natural way to divide our dataset groups, using a hierarchical clustering algorithm. Figure 2d shows the 4 vectors obtained averaging $\boldsymbol{u}$ over all $\mathcal{X}$ in the same class. These represent *types* of activity vectors. We observe that the main (but not only) difference is between households in which children (yellow dots and blue diamonds) or adults (thin purple diamonds and red squares) are more active. Figure 2e shows the corresponding plot obtained averaging $\boldsymbol{\eta}$ over all $\mathcal{X}$ belonging to the same group (as in 2d). Comparing 2e with 2d we verify if changes in the activity vector $\boldsymbol{u}$ (that explains the HCM) are related to changes in the behavior. From a visual inspection, we indeed observe that a high η_a results in a high activity u_a.

To make this observation more quantitative, for each measurement $\boldsymbol{x} \in \{\boldsymbol{u}, \boldsymbol{\eta}\}$, we write the value corresponding to age a and class p as

$$x_{a,p} = \bar{x}_a + \delta_{a,p}^{(x)}, \tag{2}$$

where $\bar{x}_a$ is the average over the $k = 4$ groups, and $\delta_{a,p}^{(x)}$ are the fluctuations. Figures 2b shows the scatter plot of the fluctuations of $\boldsymbol{\delta}^{(u)}$ and $\boldsymbol{\delta}^{(\eta)}$, evidencing a strong correlation with a highly significant (p-values less than 10^{-3}) Pearson coefficient of 0.85.

This analysis suggests that the measured contact matrix can be estimated with a high precision from aggregated (hence more easily collectable) data being the group average number of contacts per individual. Although this measure appears very efficient, it still relies on the concept of contacts. We now introduce a further parameter $\boldsymbol{y}$ that is even more easily observable than $\boldsymbol{\eta}$ and has a weaker but still strong correlation with $\boldsymbol{u}$. Specifically, the entries of $\boldsymbol{y}^{(\mathcal{X})} \in [0,1]^{n_{\text{age}}}$ indicate the fraction of people having a major activity outside the house for each age group. This quantity is expected to be negatively correlated with $\boldsymbol{u}$, since lower activities should be observed when people spend more time outside the household. Repeating the same

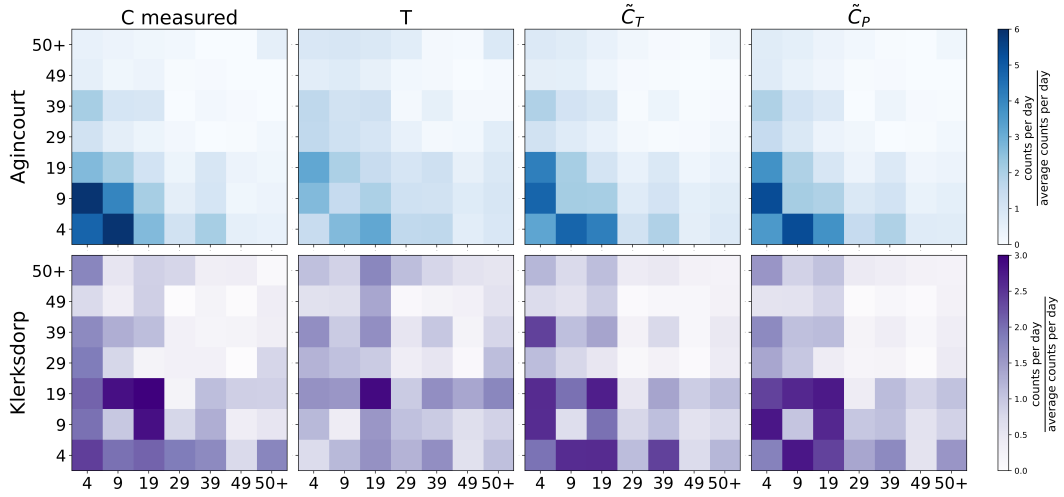

Figure 3: **Measured vs estimated normalized contact matrices in the two sites**. The first row, in blue, corresponds to Agincourt, the rural site, while the second, in purple, to Klerksdorp, the urban site. The first column shows the matrix C_{counts} aggregated over the three deployments, as measured by the proximity sensors. The second column is the corresponding random encounter matrix T. The third and the fourth are the estimate obtained by our first and second order models, respectively. All matrices are normalized by the empirical average of their entries.

procedure detailed for $\boldsymbol{\eta}$, we obtain Figures 2f showing indeed that high values of u_a are obtained for low y_a, as expected (see the red squares for the group $[40-49]$). The correlation between the fluctuations of $\boldsymbol{u}$ and $\boldsymbol{y}$ is then reported in Figure 2c, reaching a significant Pearson coefficient of -0.65.

We would like to underline that $\boldsymbol{y}$ is a very coarse grained measurement indicating whether or not each individual has a major activity outside the house and it does not include any concept of contacts.

We now discuss a refined model (with respect to Equation (1)) that keeps simultaneously the activity and the time spent at home into account. We show that this model produces better estimates of the contact matrices and can be conveniently used to predict the behavior of the *(household-deployment)* pairs originally excluded from our study.

2.3 A second order model for household interaction

We generalize the model of Equation (1) replacing $T^{(\mathcal{X})}$, the contact matrix of random encounters, which implicitly assumes that all members are simultaneously inside the household. We observed, however, that the presence of a major activity outside the household significantly impacts the HCM. To mitigate this effect we introduce the matrix $P^{(\mathcal{X})}$ that replaces $T^{(\mathcal{X})}$ in Equation (1) and that is formally defined in Section 4.1.

In $T^{(\mathcal{X})}$ we count the number of possible contacts per age pairs. In $P^{(\mathcal{X})}$, instead, each possible contact is weighted by the similarity between the individuals' contact profiles: if i only interacts in the morning and j only in the afternoon, then there is no chance that they get in contact.

The blue curve of Figure 2a compares the histogram of the cosine similarity between the actual and estimated contact matrices obtained using $P^{(\mathcal{X})}$, denoted with $\tilde{C}_P^{(X)}$. A clear gain in precision is obtained for the second order model for which the cosine similarity is greater than 0.9 for 75% of the data.

	T	$\tilde{C}_T$	$\tilde{C}_P$
Agincourt	0.83	0.95	0.96
Klerksdorp	0.89	0.95	0.98

Table 2: **Goodness of the contact matrix estimation for different methods**. The score is reported in terms of cosine similarity and the naming is consistent with Figure 3 which this table refers to.

We finally test the goodness of our model for the two sites separately on all *(household-deployment)* valid pairs, hence also those that were initially excluded because of quality issues in some (but not all) deployments. We use as $\boldsymbol{u}^{(\mathcal{X})}$ its average realization over the 2500 samples and compare the result of the predicted matrix $T, \tilde{C}_T$ and $\tilde{C}_P$ with the measured one (Figure 3), considering the two sites separately. The cosine similarities achieved by the three approximations are reported in Table 2 provide and striking evidence of how contact matrices are approximated with high precision from few age-dependent parameters.

3 Discussion

Our result brings an empirical evidence that most of the structure of contact matrices measured with high-resolution proximity sensors can be reliably captured with a simple statistical model of behavior that relies on a low-dimensional representation of HCM.

While it comes as no surprise that a parametric generalization of the matrix T would lead to better estimates, the most important aspects of our results brought the following observations forward:

- There is a simple, context-independent model, capable to estimate HCM with high precision.
- This model is highly interpretable and we expect its parameters can be directly estimated with surveys, addressing questions such as those listed in Section 1.
- All parameters are aggregated by age group and involve the behavior of single individuals and do not depend on the age class of other members. This crucial aspect naturally reduces the number of parameters of the model, making the estimation process simpler.

On the first point, the high quality and size of the PHIRST dataset gave us unprecedent insights into the problem of HCM estimation. Backed by these empirical data, not only can we say that the proposed parametric model generally improves the estimation accuracy, but we can numerically quantify it, observing very high level of agreement with the HCM obtained with the costly high resolution measurements.

On the second point, we expect this to be one of the significant outcomes of our research. In fact, we identified some practical questions to calibrate our model, bypassing the need to use proximity sensors. Given its definition provided in Section 4.1, one can immediately obtain the matrix P from the self-reported estimated daily activity patterns inside the house. Moreover, one can consider adding higher-order corrections to this model, by accounting for the time spent in isolation or the house's structure.

The last point, finally, addresses the important requirement for surveys that the questions asked should have a simple answer. Our proposed inquires revolve around self-reporting one's behavior with respect to *all* others, making no distinction on their age.

As we just mentioned, the suggested questions constitute an example of possible ways to estimate the activity parameters and are limited to the quantities that turned out to provide a significant explanation of HCM in our experiment setting. Other metadata (such as the number of rooms in the house, the wealth status or the distinction between the rural and the urban site) could potentially be informative to explain the HCM structure, even if they were not in our analysis. The lack of significance in these data probably lies in the experiment design. The correct use of the sensors mediates the HCM measurements and the model parameters cannot disentangle sociability and the compliance to the experiment, thus potentially introducing a bias that hinders the effect of other metadata on the definition of HCM.

This being said, it must be noted that the concept of compliance is not limited to the deployment of proximity sensors. In a broad sense it intrinsically affects any type of contact measurement. When HCM are estimated with surveys, for instance, it reflects one's ability to faithfully recall contacts as well as the effort dedicated to filling the form; when field observations are adopted instead, the limit is in the ability to take note of all contacts. For each technique, one would need to properly take the measurement biases into account and identify the relevant parameters that determine an individuals' propensity to engage in contact.

As a final remark, although our experiment allowed us to inspect only the household contacts in depth, we envision that similar conclusions can be extended to other settings, designing context-related contact matrices as done in [39].

Limitations

The main limitations of our study are related to the quality and nature of the available data.

The first concern is related to the time-dependent data collection component which we essentially neglected here. When dealing with contact matrices, it is customary to distinguish between weekdays and weekends. In our measurements, the first and third deployments were made asynchronously. After the cleaning procedure, it emerged that, as a consequence of the adoption of this measuring technique, weekdays and weekends are not evenly distributed among households and changes in the measured behavior are potentially associated with this effect. To cope with this problem, when dealing with asynchronous measurements it would be preferable to consider the same days of the week for all households.

A closely related concern is that we have considered all three deployments as equal, even though they correspond to rather different periods in the year. The data sparsity and quality did not allow us to detect any significant change, except for the duration of contact distribution shown in Figure 1c. It is nonetheless a very reasonable assumption that the contact behavior changes during the year. Our suggestion to investigate individuals' behavioral habits can easily overcome this problem, designing time-dependent expected matrices that could adapt even to diverse scenarios such as, for instance, during a quarantine.

On a more technical level, we reserve comment on the definition of the matrix P that is based on the similarity between the *contact* profiles measured by the sensors and that are not an easily observable quantity. An alternative, more appropriate definition would compare the similarity of the activity profiles as we already mentioned in the last section, suggesting that if two people spend simultaneously a lot of time inside the household, then it is likely that they interact. Unfortunately we do not have this kind of precise data, related to the same period of the sensors' measurements. Consequently, we used the similarity of the contacts profile as a *proxy*.

We would like to stress that, although there is a relation between the number of contacts and the similarity of the contact profiles, these two are not directly related. In fact, to build P, time is coarse-grained at a hour level and all the days of the deployment are merged. As a consequence, for each individual we have a binary indicator on whether or not he/she interacted with someone in a particular hour of the day during the deployment. We therefore believe that P (or a slightly modified version of it) can accurately be estimated from simple collectable data, as claimed through the paper.

Conclusion

These limitations being accounted for, our study poses solid and methodological observations to characterize HCM. Even if the setting considered is limited to a particular context, the simplicity and interpretability of our proposed model suggests that it has the potential of being easily generalized. As a practical application, our results can impact the strategy to design the surveys currently adopted to quantify social contacts to mitigate the Covid19 pandemic [41, 42].

4 Methods

All experiments were performed in accordance with relevant guidelines and regulations: ethics permission to conduct the experiment was received from the Wits Human Research Ethics Committee (Medical) (ethics reference no. 150808) as well as the Mpumalanga Provincial Research & Ethics Committee. Written informed consent was sought and received from all participants or their caregivers.

Let us now more rigorously define the HCM our main result relies upon. These matrices are obtained from individuals high spatio-temporal resolution proximity measurements.

4.1 Contact matrices definitions

We here provide a formal definition of all contact matrices appearing in the main text. These are all square, symmetric matrices of size n_{age}, the number of age bins considered. Let (h, d) be a household deployment pair, then for a pair of age groups a, b

$$\left(C^{(h,d)}_{\text{counts}}\right)_{ab} = \text{ number of contacts per day between } a \text{ and } b$$
$$\left(C^{(h,d)}_{\text{sec}}\right)_{ab} = \text{ time of interaction per day between } a \text{ and } b.$$

These matrices should be compared with their "expectation", or, more precisely, with the contact matrix obtained assuming a given household line-up and that people interact at random. This is given by [39]:

$$T^{(h,d)}_{ab} = \frac{\Phi^{(h,d)}_a(\Phi^{(h,d)}_b - \delta_{ab})}{\rho^{(h,d)} - 1},$$

where $\Phi^{(h,d)}_a$ is the number of people in the age group a in household h and deployment d; $\rho^{(h,d)} = \sum_a \Phi^{(h,d)}$ is the total number of people in the household in that deployment.

As an extension of $T^{(h,d)}$, we further introduce the matrix $P^{(h,d)}$. Let $\boldsymbol{v}_i \in \{0,1\}^{24}$ be a Boolean activity vector related to the individual i, denoting the presence of a contact for each hour of the day. The definition of $P^{(h,d)}$ then reads

$$P^{(h,d)}_{ab} = \frac{1}{\rho^{(h,d)} - 1} \sum_{i \in V^{(h,d)}_a} \sum_{j \in V^{(h,d)}_b \setminus \{i\}} \frac{\boldsymbol{v}_i^T \boldsymbol{v}_j}{24},$$

where $V^{(h,d)}_a$ is the set of all individuals in the age group a of the pair (h,d). Note that if $\boldsymbol{v}_i = \mathbf{1}_{24}$ for all i, the definition of $P^{(h,d)}$ corresponds to the one of $T^{(h,d)}$. The scalar product between $\boldsymbol{v}_i^T \boldsymbol{v}_j$ quantifies the time in which i and j had *simultaneously* contacts with members inside the household. If it equals zero, then there is no chance that i and j got in contact at all. In other words, the matrix $P^{(h,d)}$ predicts the contact rate assuming people get in proximity at random, but keeping into account that people are not always and simultaneously inside the house.

Letting M be any of the above matrices $C_{\text{counts}}, C_{\text{sec}}, T, P$, we define $M^{(\mathcal{X})}$ for a set $\mathcal{X}$ of *(household-deployment)* pairs as the empirical average of $M^{(x)}$ for all $x \in \mathcal{X}$. With this notation at hand, we finally introduce the matrices $R^{(\mathcal{X})}_{\text{counts/sec}}$ as

$$\left(R^{(\mathcal{X})}_{\text{counts/sec}}\right)_{ab} = \begin{cases} \gamma^{(\mathcal{X})} \frac{\left(C^{(\mathcal{X})}_{\text{counts/sec}}\right)_{ab}}{T^{(\mathcal{X})}_{ab}} & \text{if } T^{(\mathcal{X})}_{ab} \neq 0 \\ 1 & \text{else,} \end{cases}$$

where $\gamma^{(\mathcal{X})}$ is a constant to impose that the average of $R^{(\mathcal{X})}$ equals one. In words, the entries of $R^{(\mathcal{X})}$ exceed one for the pairs that interact more than expected and are below one otherwise. If a pair cannot have interactions ($T^{(\mathcal{X})}_{ab} = 0$), then we conventionally set $R^{(\mathcal{X})} = 1$.

4.2 Model test pipeline

We here detail and motivate the procedure adopted in Section 2.2 to evaluate the role of the parameter $\boldsymbol{u}^{(\mathcal{X})}$.

Sampling the villages

First of all, our goal assumes that we can distinguish different types of contact matrices, given a context created by metadata. In order to attain this objective it is necessary to define a consistent distance to compare contact matrices. When comparing different households, however, one has to consider that typically there are some age groups with no individuals. More formally, this means that for some age group a, $\Phi^{(h,d)}_a = 0$. In the extreme case in which we compare two orthogonal *(household-deployment)* pairs for which

$(\mathbf{\Phi}^{(h,d)})^T \mathbf{\Phi}^{(h',d')} = 0$ there is no way to consistently compare (h,d) with (h',d') because the corresponding contact matrices are complementary, *i.e.* the zeros one correspond to the non-zeros of the other.

To address this problem, we choose to compare *villages* $\mathcal{X}$, *i.e.* small groups of (h,d) pairs that guarantee that $\Phi_a^{(\mathcal{X})} > 0$ for all a. To build the samples $\mathcal{X}$ we then first select some (h,d) at random, with the constraint to achieve $\Phi_a^{(\mathcal{X})} > 0$ for all a (we sample only the pairs that can contribute to increasing the zeros entries of this vector) and then we randomly pick other pairs until the fixed size of $\mathcal{X}$ is reached. We set $|\mathcal{X}| = 8$ because it is heuristically a good trade-off between two competing effects: if $|\mathcal{X}|$ is too low there is a possibility of over-representing households with elderly members that are fewer and hence more valuable to get the condition $\Phi_a^{(\mathcal{X})} > 0$ for all a; on the other hand, very large values of $|\mathcal{X}|$ will tend towards an "averaging" effect that leads all *villages* to be very similar to one-another.

Optimization process

Given the samples of $\{\mathcal{X}\}$ we now compute the value $\boldsymbol{u}^{(\mathcal{X})}$ as the result of the following optimization problem

$$\boldsymbol{u}^{(\mathcal{X})} = \underset{\boldsymbol{v} \,:\, \boldsymbol{v}^T\mathbf{1}=\text{const}}{\arg\min} \; d_{\text{C}} \left(C_{\text{counts}}^{(\mathcal{X})}, T^{(\mathcal{X})} \circ \boldsymbol{v}\boldsymbol{v}^T \right),$$

where '$\circ$' denotes the entry-wise Hadamard product and d_{C} is a modified Canberra distances. Let A, B be two symmetric matrices of size n_{age}, then

$$d_{\text{C}}(A,B) = \sum_{i=1}^{n_{\text{age}}} \sum_{j \le i} \frac{|\tilde{A}_{ij} - \tilde{B}_{ij}|}{|\tilde{A}_{ij}| + |\tilde{B}_{ij}|},$$

where $\tilde{M}$ is the matrix M divided by its mean. This choice of the distance is motivated by the two following points.

1. The entries of $C_{\text{counts}}^{(\mathcal{X})}$ range in different orders of magnitude as shown in Figure 3. The cosine similarity is meaningful to quantify the proximity of two matrices but it naturally tends to give more weight to entries with a larger magnitude. When used inside an optimization (and not a single evaluation), it may lead to bad estimates of the small entries of $C_{\text{counts}}^{(\mathcal{X})}$. This problem is solved by introducing a relative distance such as d_{C}, which is unfit to evaluate single evaluations because it can be largely affected by the fluctuations related to small entries.

2. The modified Canberra distance compares a normalized version of the contact matrices because we are interested in comparing their *structures* rather than their intensities. We then have for any $\alpha, \beta > 0$, $d_{\text{C}}(\alpha A, \beta B) = d_{\text{C}}(A,B)$.

Occupation parameter

We here detail the strategy to determine the vectors $\boldsymbol{y}, \boldsymbol{\eta}$ appearing in Figure 2b, c, referred to as *occupation* and *compliance* vector respectively.

In the PHIRST data collection process, the participants were asked to specify locations or activities in which they spend more than three hours a day for more than three days per week. The options to choose from included: school, university, work, pub, social clubs, hanging out with friends, street vendors and church. We then define a Boolean variable for each person indicating whether or not he/she has a major activity outside the household, *i.e.* if he/she answered positively to *any* of the questions above. The value of $y_a^{(\mathcal{X})}$ is the average of the Boolean indicator for all people of age a in $\mathcal{X}$.

Data availability

Along with this paper, the contact matrices aggregated at the household level are publicly shared at github.com/lorenzodallamico/PHIRST_CM.

References

[1] Jacco Wallinga, W John Edmunds, and Mirjam Kretzschmar. Perspective: human contact patterns and the spread of airborne infectious diseases. *TRENDS in Microbiology*, 7(9):372–377, 1999.

[2] Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

[3] Lauren Meyers. Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bulletin of the American Mathematical Society*, 44(1):63–86, 2007.

[4] Robert Verity, Lucy C Okell, Ilaria Dorigatti, Peter Winskill, Charles Whittaker, Natsuko Imai, Gina Cuomo-Dannenburg, Hayley Thompson, Patrick GT Walker, Han Fu, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet infectious diseases*, 20(6):669–677, 2020.

[5] Patrick GT Walker, Charles Whittaker, Oliver J Watson, Marc Baguelin, Peter Winskill, Arran Hamlet, Bimandra A Djafaara, Zulma Cucunubá, Daniela Olivera Mesa, Will Green, et al. The impact of covid-19 and strategies for mitigation and suppression in low-and middle-income countries. *Science*, 369(6502):413–422, 2020.

[6] Kaiyuan Sun, Wei Wang, Lidong Gao, Yan Wang, Kaiwei Luo, Lingshuang Ren, Zhifei Zhan, Xinghui Chen, Shanlu Zhao, Yiwei Huang, et al. Transmission heterogeneities, kinetics, and controllability of sars-cov-2. *Science*, 371(6526):eabe2424, 2021.

[7] T House and MJ Keeling. Household structure and infectious disease transmission. *Epidemiology & Infection*, 137(5):654–661, 2009.

[8] Nele Goeyvaerts, Eva Santermans, Gail Potter, Andrea Torneri, Kim Van Kerckhove, Lander Willem, Marc Aerts, Philippe Beutels, and Niel Hens. Household members do not contact each other at random: implications for infectious disease modelling. *Proceedings of the Royal Society B*, 285(1893):20182201, 2018.

[9] Zachary McCarthy, Yanyu Xiao, Francesca Scarabel, Biao Tang, Nicola Luigi Bragazzi, Kyeongah Nah, Jane M Heffernan, Ali Asgary, V Kumar Murty, Nicholas H Ogden, et al. Quantifying the shift in social contact patterns in response to non-pharmaceutical interventions. *Journal of Mathematics in Industry*, 10(1):1–25, 2020.

[10] Giulia Cencetti, Gabriele Santin, Antonio Longa, Emanuele Pigani, Alain Barrat, Ciro Cattuto, Sune Lehmann, Marcel Salathe, and Bruno Lepri. Digital proximity tracing on empirical contact networks for pandemic control. *Nature communications*, 12(1):1–12, 2021.

[11] Jacco Wallinga, Peter Teunis, and Mirjam Kretzschmar. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *American journal of epidemiology*, 164(10):936–944, 2006.

[12] Joe Hilton and Matt J Keeling. Incorporating household structure and demography into models of endemic disease. *Journal of the Royal Society Interface*, 16(157):20190317, 2019.

[13] Wei Li, Bo Zhang, Jianhua Lu, Shihua Liu, Zhiqiang Chang, Cao Peng, Xinghua Liu, Peng Zhang, Yan Ling, Kaixiong Tao, et al. Characteristics of household transmission of covid-19. *Clinical Infectious Diseases*, 71(8):1943–1946, 2020.

[14] Simon P Johnstone-Robertson, Daniella Mark, Carl Morrow, Keren Middelkoop, Melika Chiswell, Lisa DH Aquino, Linda-Gail Bekker, and Robin Wood. Social mixing patterns within a south african township community: implications for respiratory disease transmission and control. *American journal of epidemiology*, 174(11):1246–1255, 2011.

[15] Moses Chapa Kiti, Timothy Muiruri Kinyanjui, Dorothy Chelagat Koech, Patrick Kiio Munywoki, Graham Francis Medley, and David James Nokes. Quantifying age-related rates of social contact using diaries in a rural coastal population of kenya. *PloS one*, 9(8):e104786, 2014.

[16] O le Polain de Waroux, Sandra Cohuet, Donny Ndazima, AJ Kucharski, Aitana Juan-Giner, Stefan Flasche, Elioda Tumwesigye, Rinah Arinaitwe, Juliet Mwanga-Amumpaire, Yap Boum, et al. Characteristics of human encounters and social mixing patterns relevant to infectious diseases spread by close contact: a survey in southwest uganda. *BMC infectious diseases*, 18(1):1–12, 2018.

[17] Deus Thindwa, Kondwani C Jambo, John Ojal, Peter MacPherson, Mphatso Dennis Phiri, Amy Pinsent, McEwen Khundi, Lingstone Chiume, Katherine E Gallagher, Robert S Heyderman, et al. Social mixing patterns relevant to infectious diseases spread by close contact in urban blantyre, malawi. *Epidemics*, page 100590, 2022.

[18] Mohsen Naghavi, Amanuel Alemu Abajobir, Cristiana Abbafati, Kaja M Abbas, Foad Abd-Allah, Semaw Ferede Abera, Victor Aboyans, Olatunji Adetokunboh, Ashkan Afshin, Anurag Agrawal, et al. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the global burden of disease study 2016. *The lancet*, 390(10100):1151–1210, 2017.

[19] Joël Mossong, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, Jacco Wallinga, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS medicine*, 5(3):e74, 2008.

[20] Kiesha Prem, Alex R Cook, and Mark Jit. Projecting social contact matrices in 152 countries using contact surveys and demographic data. *PLoS computational biology*, 13(9):e1005697, 2017.

[21] Dina Mistry, Maria Litvinova, Ana Pastore y Piontti, Matteo Chinazzi, Laura Fumanelli, Marcelo FC Gomes, Syed A Haque, Quan-Hui Liu, Kunpeng Mu, Xinyue Xiong, et al. Inferring high-resolution human mixing patterns for disease modeling. *Nature communications*, 12(1):1–12, 2021.

[22] Gail E Potter, Mark S Handcock, Ira M Longini Jr, and M Elizabeth Halloran. Estimating within-household contact networks from egocentric data. *The annals of applied statistics*, 5(3):1816, 2011.

[23] Timo Smieszek, Elena U Burri, Robert Scherzinger, and Roland W Scholz. Collecting close-contact social mixing data with contact diaries: reporting errors and biases. *Epidemiology & infection*, 140(4):744–752, 2012.

[24] Rossana Mastrandrea and Alain Barrat. How to estimate epidemic risk from incomplete contact diaries data? *PLoS computational biology*, 12(6):e1005002, 2016.

[25] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne Régis, Bruno Lina, et al. High-resolution measurements of face-to-face contact patterns in a primary school. *PloS one*, 6(8):e23176, 2011.

[26] Philippe Vanhems, Alain Barrat, Ciro Cattuto, Jean-François Pinton, Nagham Khanafer, Corinne Régis, Byeul-a Kim, Brigitte Comte, and Nicolas Voirin. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PloS one*, 8(9):e73970, 2013.

[27] Laura Ozella, Francesco Gesualdo, Michele Tizzoni, Caterina Rizzo, Elisabetta Pandolfi, Ilaria Campagna, Alberto Eugenio Tozzi, and Ciro Cattuto. Close encounters between infants and household members measured through wearable proximity sensors. *PloS one*, 13(6):e0198733, 2018.

[28] Michele Starnini, Bruno Lepri, Andrea Baronchelli, Alain Barrat, Ciro Cattuto, and Romualdo Pastor-Satorras. Robust modeling of human contact networks across different scales and proximity-sensing techniques. In *International Conference on Social Informatics*, pages 536–551. Springer, 2017.

[29] Moses Chapa Kiti, Alessia Melegaro, Ciro Cattuto, and David James Nokes. Study design and protocol for investigating social network patterns in rural and urban schools and households in a coastal setting in kenya using wearable proximity sensors. *Wellcome open research*, 4, 2019.

[30] Cheryl Cohen, Meredith L McMorrow, Neil A Martinson, Kathleen Kahn, Florette K Treurnicht, Jocelyn Moyes, Thulisa Mkhencele, Orienka Hellferscee, Limakatso Lebina, Matebejane Moroe, et al. Cohort profile: A prospective household cohort study of influenza, respiratory syncytial virus and other respiratory pathogens community burden and transmission dynamics in south africa, 2016–2018. *Influenza and Other Respiratory Viruses*, 15(6):789–803, 2021.

[31] Jackie Kleynhans, Stefano Tempia, Meredith L McMorrow, Anne von Gottberg, Neil A Martinson, Kathleen Kahn, Jocelyn Moyes, Thulisa Mkhencele, Limakatso Lebina, F Xavier Gómez-Olivé, et al. A cross-sectional study measuring contact patterns using diaries in an urban and a rural community in south africa, 2018. *BMC public health*, 21(1):1–10, 2021.

[32] Cheryl Cohen, Jackie Kleynhans, Anne von Gottberg, Meredith L McMorrow, Nicole Wolter, Jinal N Bhiman, Jocelyn Moyes, Mignon du Plessis, Maimuna Carrim, Amelia Buys, et al. Sars-cov-2 incidence, transmission and reinfection in a rural and an urban setting: results of the phirst-c cohort study, south africa, 2020–2021. *Medrxiv*, 2021.

[33] Jackie Kleynhans, Stefano Tempia, Nicole Wolter, Anne von Gottberg, Jinal N Bhiman, Amelia Buys, Jocelyn Moyes, Meredith L McMorrow, Kathleen Kahn, F Xavier Gómez-Olivé, et al. Sars-cov-2 seroprevalence in a rural and urban household cohort during first and second waves of infections, south africa, july 2020–march 2021. *Emerging infectious diseases*, 27(12):3020, 2021.

[34] Cheryl Cohen, Jackie Kleynhans, Jocelyn Moyes, Meredith L McMorrow, Florette K Treurnicht, Orienka Hellferscee, Azwifarwi Mathunjwa, Anne von Gottberg, Nicole Wolter, Neil A Martinson, et al. Asymptomatic transmission and high community burden of seasonal influenza in an urban and a rural community in south africa, 2017–18 (phirst): a population cohort study. *The Lancet Global Health*, 9(6):e863–e874, 2021.

[35] Deus Thindwa, Nicole Wolter, Amy Pinsent, Maimuna Carrim, John Ojal, Stefano Tempia, Jocelyn Moyes, Meredith McMorrow, Jackie Kleynhans, Anne von Gottberg, et al. Estimating the contribution of hiv-infected adults to household pneumococcal transmission in south africa, 2016–2018: A hidden markov modelling study. *PLoS computational biology*, 17(12):e1009680, 2021.

[36] Eduan Wilkinson, Marta Giovanetti, Houriiyah Tegally, James E San, Richard Lessells, Diego Cuadros, Darren P Martin, David A Rasmussen, Abdel-Rahman N Zekri, Abdoul K Sangare, et al. A year of genomic surveillance reveals how the sars-cov-2 pandemic unfolded in africa. *Science*, 374(6566):423–431, 2021.

[37] Ledor S Igboh, Meredith Mcmorrow, Stefano Tempia, Gideon O Emukule, Ndahwouh Talla Nzussouo, Margaret Mccarron, Thelma Williams, Vashonia Weatherspoon, Ann Moen, Derrar Fawzi, et al. Influenza surveillance capacity improvements in africa during 2011-2017. *Influenza and other respiratory viruses*, 15(4):495–505, 2021.

[38] Stefano Tempia, Sibongile Walaza, Jinal N Bhiman, Meredith L McMorrow, Jocelyn Moyes, Thulisa Mkhencele, Susan Meiring, Vanessa Quan, Kate Bishop, Johanna M McAnerney, et al. Decline of influenza and respiratory syncytial virus detection in facility-based surveillance during the covid-19 pandemic, south africa, january to october 2020. *Eurosurveillance*, 26(29):2001600, 2021.

[39] Laura Fumanelli, Marco Ajelli, Piero Manfredi, Alessandro Vespignani, and Stefano Merler. Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread. 2012.

[40] Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[41] Frederik Verelst, Lisa Hermans, Sarah Vercruysse, Amy Gimma, Pietro Coletti, Jantien A Backer, Kerry LM Wong, James Wambua, Kevin van Zandvoort, Lander Willem, et al. Socrates-comix: a platform for timely and open-source contact mixing data during and in between covid-19 surges and interventions in over 20 european countries. *BMC medicine*, 19(1):1–7, 2021.

[42] Carl E Koppeschaar, Vittoria Colizza, Caroline Guerrisi, Clément Turbelin, Jim Duggan, W John Edmunds, Charlotte Kjelsø, Ricardo Mexia, Yamir Moreno, Sandro Meloni, et al. Influenzanet: citizens among 10 countries collaborating to monitor influenza in europe. *JMIR public health and surveillance*, 3(3):e7429, 2017.

[43] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PloS one*, 5(7):e11596, 2010.

Author contributions

CCa, CCo, ST conceived and designed the study. LD conceived the model, performed the numerical simulations, the statistical analysis and wrote the manuscript. JK detailed the collection procedure in the supplementary material. JK, MM, NW, BL, RGW, CCo, ST led the data collection. LD, LG, MT, CCa analyzed the data. All authors read and approved the final manuscript.

Funding

This work was supported by the National Institute for Communicable Diseases of the National Health Laboratory Service and the U.S. Centers for Disease Control and Prevention [co-operative agreement number: 1U01IP001048]. The funders had no role in design, analysis or interpretation of data. LD and CCa acknowledge support from the Lagrange Project of ISI Foundation funded by CRT Foundation, and from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101016233 (PERISCOPE). LG and MT acknowledge support from the Lagrange Project of ISI Foundation funded by CRT Foundation.

Competing interests

CCo has received grant support from Sanofi Pasteur, US CDC, Welcome Trust, Programme for Applied Technologies in Health (PATH), Bill & Melinda Gates Foundation and South African Medical Research Council (SA-MRC). NW reports receiving grants from Sanofi Pasteur, US CDC and the Bill & Melinda Gates Foundation. All other authors do not report any competing interests.

Supplementary information

S.1 Data collection

Proximity sensors

Proximity data are measured with the `SocioPatterns` sensors that we here succinctly introduce, addressing the interested reader to [43] for a more detailed reference.

Their functioning is based on the emission of low-power and low-frequency signals. Participants are asked to wear the sensor on their chest, so that when they engage in a face-to-face interaction with another participant, the respective sensors can exchange packets of information with a frequency that does not exceed one packet per second. A contact is measured if, in the time-span of 20 seconds, two sensors exchange at least one packet, recording the unique identifier of the interacting sensor, the time at which the interaction occurred and the attenuation of the signal from the sender to the receiver. This attenuation is directly connected to the distance between the two sensors and can be used to filter only close proximity interactions.

Additionally, each sensor periodically records some status properties that allow the user to know whether it is functioning correctly or not. Among these, an accelerometer allows one to know if the sensor is moving (hence if it is worn) or not.

Collection procedure

The PHIRST study was a prospective household cohort study described previously [30, 34] In short, we enrolled a new cohort of households in 2016, 2017 and 2018 at two sites in South Africa (urban: Klerksdorp, North West and rural: Agincourt, Mpumalanga) and followed households up for 8 to 10 months. Consenting household members were visited twice weekly to collect nasopharyngeal specimens and self-reported symptom data to investigate the transmission of influenza respiratory syncytial virus (RSV) and Streptococcus pneumoniae. In the 2018 cohort, we deployed wearable proximity sensors for 10 – 14 days to all consenting household members to measure high-resolution household contact patterns during three periods of the year. Sensors were worn in PVC pouches on the chest or on a lanyard. Participants were requested to put the sensor on in the morning, keep it on the entire day (even when leaving the home), take it off at night and store it separately from other household member's sensors. Not all participants felt comfortable wearing sensors outside of the home and instead took sensors off when not at home. Participants were requested to complete a log to indicate the times the sensor was put on and taken off during the day. During the twice weekly visits to the household, study staff reminded participants to wear the sensors, monitored if all sensors were still working, and replaced batteries where sensors had stopped working. After at least a ten-day deployment, sensors were collected at the next routine household visit of study staff to the household and taken back to the study office where batteries were removed.

Data cleaning

The data cleaning procedure is described as follows:

1. All contacts measured by non-moving sensors are removed: this is to avoid including spurious contacts between sensors that are, for instance, kept in proximity inside a drawer.

2. Contacts are filtered according to their attenuation and only those with an attenuation of $-70\ dBm$ or less are kept. This threshold corresponds to an interaction between two sensors that are approximately at 2 meters, even if this is a context-dependent relation that depends on external parameters, such as, for instance, humidity.

3. All contacts happening before the official beginning of the deployment and after its end are removed. These contacts may exist, because sensors may be collected on different dates from the ones of the planned experiment, but they are removed because sensors' use may be non-systematic, hence unreliable. Moreover, the first and last day of measurement are removed as well. During these days, very intense activity patterns are typically observed due to the interaction with the people dispatching the sensors. Since this kind of interaction deviates from the standard conditions, it is not considered.

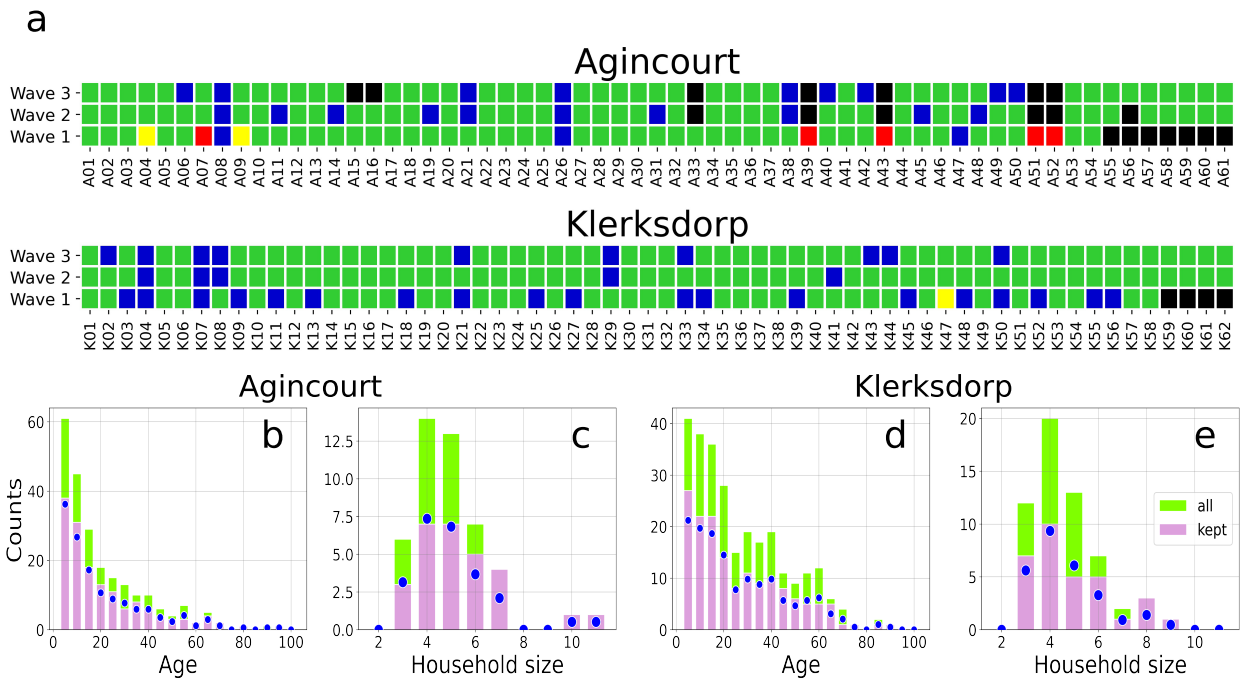

Figure S.1: **Raw data characteristics**. **a**: *data quality*. For each *(household-deployment)* we assign a color code: black indicates that the household did not participate; red that all household's sensors had data quality issues and did not provide valid measurements; blue that there are less than two days of measurement; yellow that a non circadian activity is observed; green none of the above. **b** and **d**: *age distribution in Agincourt and Klerksdorp, respectively*. The green bars are referred to the whole data-set, while the purple one only to the 60 households with valid measurements in all three deployments (see *a*). Blue dots are obtained multiplying the height of the green bars for the fraction of the included households. **c** and **e**: *household size distribution*. Legends and colors follow the description of *b* and *d*.

4. The data collected by the sensors contains information on the hardware identification code. A mapping relates this identifier with the individuals' anonymous identification code that allows us to relate contacts and meta-data. Errors at this stage make it impossible to relate contacts to people and results in the red dots shown in Figure S.1a.

5. As a minimal request, we impose that, after this cleaning procedure, a deployment can be considered valid only if it has two or more days of measurement. We found this to be a good trade-off between high quality data to work with and a still rather comprehensive inclusion principle. Household-deployment pairs that do not fulfill this condition are denoted in blue in Figure S.1a.

6. Finally, non-circadian activity patterns are identified. A high activity during the night was observed in only three cases (yellow dots of Figure S.1a) during the first deployment and are likely symptomatic of a misuse of the sensors happened at the beginning of the experiment.

Only the households in which all three deployments led to valid measurements (all green dots in Figure S.1a) were included in our study. Figures S.1b, c, d, e further show the age and household size histograms for the whole dataset against its cleaned version, showing that our inclusion principle did not affect either of the four distributions.