Estimating household contact matrices structure from easily collectable metadata

Lorenzo Dall'Amico[1,*] Jackie Kleynhans[2,3] Laetitia Gauvin[1,4] Michele Tizzoni[1,5] Laura Ozella[1] Mvuyo Makhasi[2] Nicole Wolter[2,6] Brigitte Language[7] Ryan G. Wagner[8] Cheryl Cohen[2,3] Stefano Tempia[2,3] Ciro Cattuto[1,9]

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 Institute for Research on sustainable Development, UMR215 PRODIG, Aubervilliers, France
5 Department of Sociology and Social Research, University of Trento, Trento, Italy
6 School of Pathology, University of the Witwatersrand, Johannesburg, South Africa
7 Unit for Environmental Science and Management, Climatology Research Group, North-West University, Potchefstroom, South Africa
8 MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt)
9 University of Turin, Department of Informatics, Turin, 10124, Italy
***** Corresponding to: `lorenzo.dallamico@isi.it`

Abstract

Contact matrices are a commonly adopted data representation, used to develop compartmental models for epidemic spreading, accounting for the contact heterogeneities across age groups. Their estimation, however, is generally time and effort consuming and model-driven strategies to quantify the contacts are often needed. In this article we focus on household contact matrices, describing the contacts among the members of a family and develop a parametric model to describe them. This model combines demographic and easily quantifiable survey-based data and is tested on high resolution proximity data collected in two sites 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 transmitted through close proximity contacts [1] and the modeling thereof is a problem of great 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 [2–6]. 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 [8–10]. 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 [11–15].

Contact matrices are generally estimated through surveys in which the participants have to self-report their contacts in terms of number, duration and (presumed) age of the interacting individual [15–18]. Known limitations of this technique include under-reporting of contacts and overestimation of their durations [19, 20]. Determining *household contact matrices* (HCM) is resource-intensive, hardly scalable and technically challenging, especially in low-resource sub-Saharan African countries with high infectious diseases burden and where the data collection is still very limited [21–25]. Consequently, a growing attention is devoted to theoretically model HCM. Some of the most popular models to estimate contact matrices rely on the demographic properties of the population under study [17, 26], eventually taking the setting (*e.g.* school, work, home) in which the interactions take place into account. These models assume that the number of contacts between age groups approximately scales as the product of the two population sizes involved, *i.e.* the number of all possible pairs. In [16] the authors further considered how to make estimates of contact matrices available in countries where the mixing patterns were not directly estimated. More recently, [27] introduced generalized contact matrices in which socio-economic factors are included as well. The authors propose a simple model inducing assortative mixing that is pervasively observed in real-world data.

Here we consider HCM obtained from proximity sensors, encoding the sequence of contacts among a group of selected participants with high resolution in space and time. The proximity sensors are developed by the `SocioPatterns` collaboration (sociopatterns.org, [28]) and allow us to study and model human dynamics [21, 29–33] and directly estimate HCM by aggregating individuals' contacts across time. We analyze the data collected during the PHIRST study [34, 35], a 3-year long experiment conducted in South Africa, designed to provide reliable data-driven guidance to limit viral transmission [34, 36–42]. We show that, although demographic properties are determinant in shaping the HCM, they are insufficient to accurately capture the contacts structure and further age-dependent parameters must be introduced to model the higher sociability typically observed among young people [43]. Our parametric model can be calibrated with surveys but, unlike the direct estimation of the full contact matrix, they introduce several advantages. Firstly one only needs to report one's age and not the age of the other interacting individuals, making the estimation process more reliable by design. Secondly, the number of parameters to be estimated scales linearly with the number of age bins (and not quadratically) and the binning itself can be chosen a posteriori. Our method can thus be seen as a reliable compromise between a parameter-free demographic model and a direct estimation of the contact matrix from surveys. Testing our results on the high-resolution measurements, we show that one can approximate the HCM with a cosine similarity equal to 0.96 and 0.98 in the two sites.

2 Data descriptive statistics

We now provide an overview of the data collection strategy, as well as some basic descriptive statistics.

2.1 Data collection

The PHIRST study was a prospective household cohort study described previously in [34, 38]. We enrolled a cohort of households 2018 at two sites in South Africa (urban: Klerksdorp, North West and rural: Agincourt, Mpumalanga) and followed households

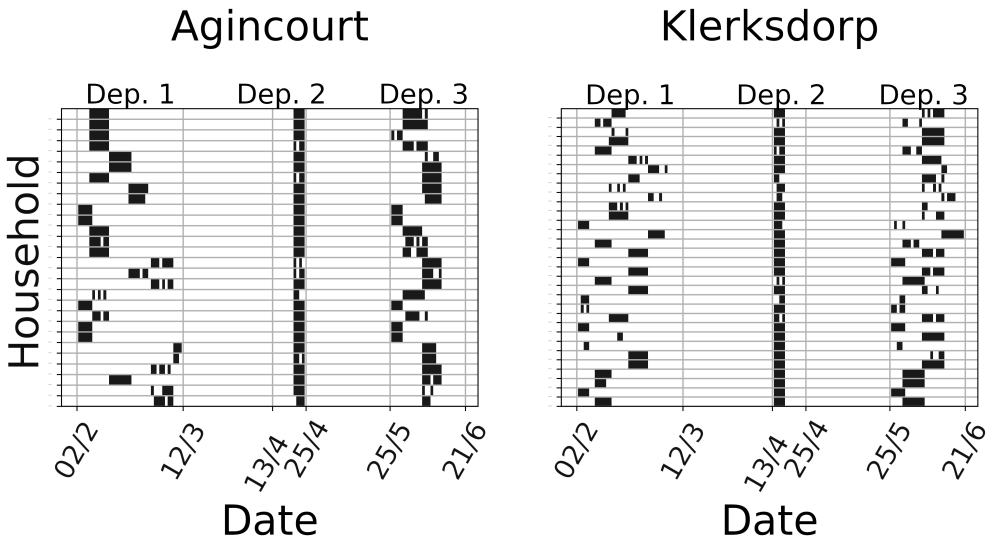

Fig 1. Data collection schedule for the 60 selected households. Each row corresponds to a household with the rural site on the left and the urban site on the right. Time is displayed on the x axis and dates are reported in the day/month format. Vertical gray lines correspond to the beginning and end of each deployment. A black dot indicates that at least one contact was measured, while a white one that no contact was recorded on that day.

up for 8 to 10 months. Wearable proximity sensors were deployed for 10 to 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 wear 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 diary to indicate the times the sensor was put on and taken off during the day. Twice a week, the staff visited each household and 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 and data was downloaded from the sensors. After the data cleaning procedure, detailed in Section S.1, our dataset is composed of 307 individuals subdivided into 60 households. For consistency, we choose to consider only households for which the data quality was sufficiently high in all three deployments. The exclusion can be due to the displacement of some individuals or to technical problems with specific sensors. As discussed in the supplementary material, the cleaned dataset is representative of the original both in terms of size and age distributions. Figure 1 summarizes the data collection schedule.

Ethical approval

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.

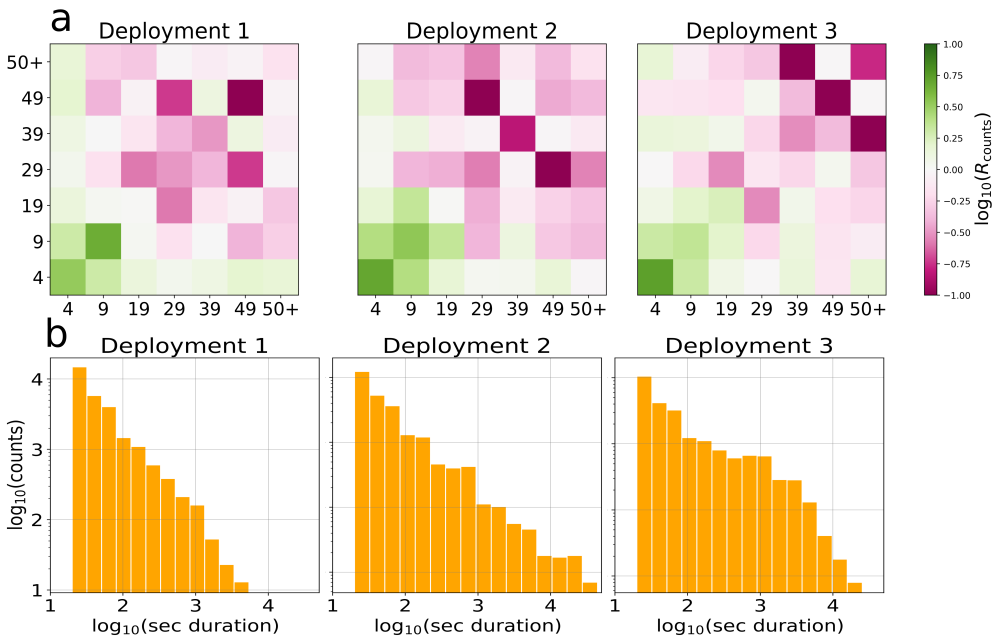

Fig 2. Properties of the measured data: **a**: normalized contact matrix across the three deployments. The color code refers to the values of the logarithm of R_{counts} whose entries are proportional to the ratio between the number of contacts and the number of possible interacting pairs, setting the mean of R_{counts} to 1. The two axis correspond to the age groups and the number reported indicates the highest age of each group. **b**: contact duration distribution expressed as number of seconds of interaction across the three deployments in logarithmic scale.

2.2 Contact matrices

In this section we describe the properties of the contact matrices as measured by the proximity sensors, after having provided some formal definitions.

Definitions

Contact matrices incorporate the contacts subdivided by age groups. They 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 because of the large proportion of population in those age brackets, shown in Figure S.1. Each HCM refers to a single household and a specific deployment. We thus consider a total of 180 HCM. With the notation C, S we refer to the contact matrices storing the counts/time of interaction between pairs of age groups respectively, or, more precisely

$$\begin{aligned} C_{ab} &= \text{ number of contacts per day between } a \text{ and } b, \\ S_{ab} &= \text{ total time in contact per day between } a \text{ and } b. \end{aligned}$$

These matrices should be compared with their expectation, *i.e.* with the contact matrix obtained assuming a given household line-up and that people interact at random. This is given by [26]:

R_C	First	Second	Third
First	1	0.94	0.89
Second	0.94	1	0.94
Third	0.89	0.94	1

R_S	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 contact matrices R_C (left) and R_S (right) in the three deployments.

$$T_{ab} = \frac{\Phi_a \Phi_b - \delta_{ab}}{\rho - 1}, \tag{1}$$

where Φ_a is the number of people in the age group a in a given HCM; $\rho = \sum_a \Phi_a$ is the total number of people and δ_{ab} is the Kroeneker delta (equal to 1 is $a = b$ and equal to 0 otherwise). For a set $\mathcal{X}$ of HCM, we define $C^{(\mathcal{X})}, S^{(\mathcal{X})}, T^{(\mathcal{X})}$ as the average of the respective matrix over all $\mathcal{X}$ and $R_C^{(\mathcal{X})}$ as

$$\left(R_C^{(\mathcal{X})}\right)_{ab} = \begin{cases} \gamma^{(\mathcal{X})} \frac{C_{ab}^{(\mathcal{X})}}{T_{ab}^{(\mathcal{X})}} & \text{if } T_{ab}^{(\mathcal{X})} \neq 0 \\ 1 & \text{else} \end{cases}$$

where $\gamma^{(\mathcal{X})}$ is a constant to impose that the average of $R_C^{(\mathcal{X})}$ equals one. In an analogous way, we define $R_S^{(\mathcal{X})}$ replacing C with S. 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, we conventionally set $R^{(\mathcal{X})} = 1$. To simplify the notation, in the remainder we drop the index $\mathcal{X}$.

Properties of the measured matrices

Given that we considered the same set of households across the three deployments, changes in the HCM structure can mainly be amenable to a seasonality effect. Table 1 precisely shows the cosine similarity between R_C (left) and R_S (right) for $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3$ being the set of all households in the three deployments. The table reports high similarity values for R_C, suggesting that the structure of the contact matrix does not vary a lot across the three deployments. Smaller values are instead obtained by R_S implying that the seasonality effect majorly involves the duration (rather than the structure) of the contacts. This observation agrees with the distribution of the individual contact durations, obtained from approximately 10^5 proximity measurements shown in Figure 2b which follows a broad distribution, as expected [44]. This distribution broadens 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 2a shows instead the matrix $\log(R_C)$ across the three deployments, evidencing that younger age groups tend to interact more, regardless of the age group they are interacting with. Based on these observations, we attempt to model the matrix C whose behavior is more predictable than S. Given the result of Table 1, the deployments are treated as three independent, equally reliable measurements of the HCMs.

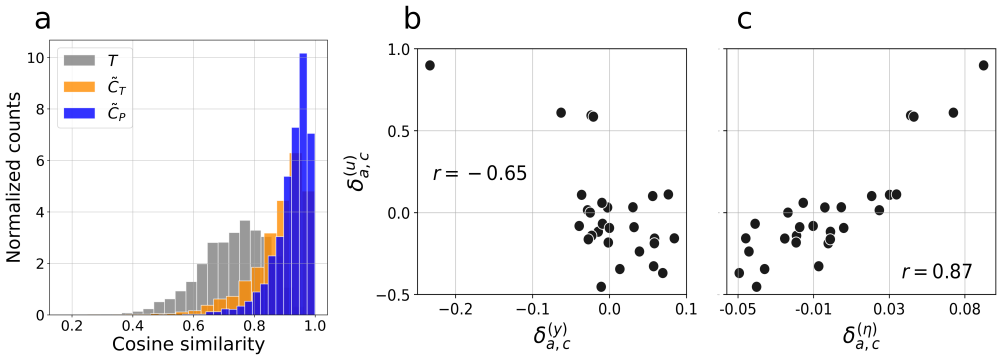

Fig 3. Test of the model for household interaction. **a**: histogram of the cosine similarity between C and its estimators. The gray curve corresponds to the histogram over the 2500 realization of $\mathcal{X}$ using T as an estimator of C. The orange curve is obtained with the first order model of (2), while the blue curve corresponds to the second order model of Section 3.2. **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 (3). The Pearson correlation coefficient r is reported in text.

3 Main result

We introduce two parametric models to approximate the HCM that combine three age-dependent parameters: the number of individuals per age group, the in-house hourly presence and an intensity of activity factor. 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^2_{age} to n_{age}.

We here propose some example of questions to estimate the in-house hourly presence and the intensity of activity factor.

- *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?*
- *How many face-to-face interactions do you have per day?*

As we will see in the remainder, these questions permit to calibrate the parameters of our model, allowing one to obtain a more faithful representation of contact matrices than the one obtained from purely demographic models. In Section 3.3 we describe some practical implications of our results and the relation to the questions listed above.

3.1 A first order model for household interaction

In this section, we define a parametric model to approximate the contact matrix C, as measured by proximity sensors. All matrices here refer to sets of HCM but we drop the index $\mathcal{X}$ to keep a light notation. Let T be the matrix defined in Equation (1). We define $\tilde{C}_T$, an approximation of C, as

$$\tilde{C}_T = T \circ \left(\boldsymbol{u}\boldsymbol{u}^T\right), \tag{2}$$

where $\boldsymbol{u} \in \mathbb{R}^{n_{\text{age}}}$ is a set of parameters that represent the activity of each age group and '$\circ$' denotes the entry-wise Hadamard product. The entries of this matrix are $(\tilde{C}_T)_{ab} = T_{ab}u_a u_b$ and a large number of interactions are expected when many members are present (large values of T_{ab}) and when they correspond to highly active age groups, such as $[0-4, 5-9]$, as per Figure 2b.

Model validation

We deploy the following steps to test our model, as detailed and motivated in Section S.2. We independently randomly sample 2500 sets $\mathcal{X}$ of 8 HCM without replacement out of the 180 available. For each sampled $\mathcal{X}$ we compute the vector $\boldsymbol{u}$ that best approximates C, minimizing a modified Canberra distance [45] between the measured and the estimated matrix, as described in Section S.2. The entries of this vector contain the activity of each age group for the set $\mathcal{X}$. Figure 3a displays the histogram of the cosine similarity between the approximation $\tilde{C}_T$ and the measured matrix C and evidences a good agreement between the two matrices with a cosine similarity equal to 0.9 or larger for 53% of the data. This similarity is of the same order of the one observed across the three deployments and reported in Table 1. Figure 3 further shows the same histogram for T being used as an estimator of C. This purely demographic model is much less accurate and reaches a cosine similarity greater than 0.9 for only 7% of the data and 50% of the data have a similarity greater or equal to 0.75.

Interpretation of the parameters

Besides the goodness of the approximation itself, our main interest is to assess whether the vector $\boldsymbol{u}$ can be estimated from easily observable quantities. To do so, for each sampled $\mathcal{X}$ we further compute the vector $\boldsymbol{\eta} \in \mathbb{R}^{n_{\text{age}}}$. Its element η_a is the number of daily interactions per individual, averaged over all individuals in a given age group a. Intuitively, $\boldsymbol{u}$ and $\boldsymbol{\eta}$ should correlate: a higher activity has to be observed when people are more active. Note that $\boldsymbol{\eta}$ aggregates *all* individual's contacts and is oblivious of the age group binning. We divide the sets $\mathcal{X}$ according to their activity vector representation $\boldsymbol{u}$ into $k = 4$ groups with a hierarchical clustering algorithm. For each $\boldsymbol{x} \in \{\boldsymbol{u}, \boldsymbol{\eta}\}$, we then write the value corresponding to age a and class c as

$$x_{a,c} = \bar{x}_a + \delta^{(x)}_{a,c}, \tag{3}$$

where $\bar{x}_a$ is the average over the 4 groups, and $\delta^{(x)}_{a,p}$ are the fluctuations. Figures 3b shows the scatter plot of the fluctuations of $\boldsymbol{\delta}^{(u)}$ and $\boldsymbol{\delta}^{(\eta)}$, evidencing a strong correlation with a highly significant (p-value 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 average number of contacts per individual in the same age group. 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 $\boldsymbol{y} \in [0,1]^{n_{\text{age}}}$ indicate the fraction of people for each age group having an occupation outside the house requiring at least three hours a day. 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 procedure detailed for $\boldsymbol{\eta}$, we obtain Figures 3f 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 reported in Figure 3c, reaching a significant Pearson coefficient of -0.65. We underline that $\boldsymbol{y}$ is a very aggregated quantity that does not directly involve contacts.

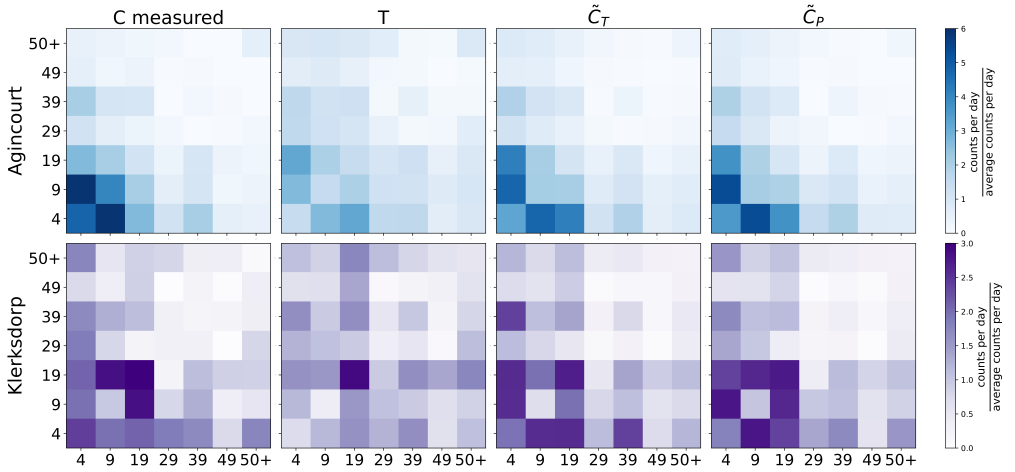

Fig 4. Measured vs estimated normalized contact matrices in the two sites. The first row, in blue, corresponds to Agincourt, the rural site, while the second, in purple, to Klerksdorp, the urban site. The first column shows the matrix C aggregated over the three deployments, as measured by the proximity sensors. The second column is the corresponding random encounter matrix T. The third and the fourth are the estimates obtained by our first and second order models, respectively. All matrices are normalized by the empirical average of their entries.

We now discuss a refined model with respect to Eq (2) that keeps simultaneously into account the activity and the time spent at home. We show that this model produces better estimates of the contact matrices and can be conveniently used to predict the HCM originally excluded from our study.

3.2 A second order model for household interaction

In Equation (1) we introduced the matrix T that encodes a purely demographic interaction model in which a higher contact rate is entirely explained by a higher number of interacting individuals. In practice, however, contacts can happen only when people are in the same physical space. To model this effect, we propose an extension of T, that we denote with P. Let $\boldsymbol{v}_i \in \{0,1\}^{24}$ be a binary-value presence vector of i, denoting the presence in the house for each hour of the day. The definition of P then reads

$$P_{ab} = \frac{1}{\rho - 1} \sum_{i \in V_a} \sum_{j \in V_b \setminus \{i\}} \frac{\boldsymbol{v}_i^T \boldsymbol{v}_j}{24} \tag{4}$$

where V_a is the set of all individuals in the age group a. Note that if $\boldsymbol{v}_{i,t} = 1$ for all i and all t, the definition of P corresponds to the one of T. 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, P 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. We generalize the model of Eq (2) replacing T with P and obtaining $\tilde{C}_P$. Practically, the proximity sensors do not provide us with the information of whether or not an individual is at home in a given moment, but only if it is interacting with another household member. For each individual we then construct a binary indicator on whether or not he/she interacted with someone in a particular hour of the day during the deployment and use this as a proxy for $\boldsymbol{v}$.

	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 4 which this table refers to.

Model testing

The blue histogram of Figure 3a shows the cosine similarity between the actual and estimated contact matrices obtained using P. A clear gain in accuracy is achieved, obtaining a cosine similarity is greater than 0.9 for 75% of the data.

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 $\boldsymbol{u}$ 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 4), considering the two sites separately. The cosine similarity scores reported in Table 2 provide and striking evidence of how contact matrices are approximated with high precision using few age-dependent parameters.

3.3 Practical implications

Let us briefly discuss some implications of our results and suggest how these could be translated into practical recommendations for data collection. Survey based estimations are, to-date, the most common and reliable way to estimate contact matrices. This method, however, has some notable limitations – that we discussed in the Section 1 – and would benefit from the design of simpler questionnaires. We highlight that one can accurately estimate HCM from self-reported quantities that are, by design, more easily and reliably estimated. Our model combines the probability that two individuals meet with an age-dependent activity driven model [46].

We suggested some examples of questions that can be formulated to calibrate our model. For instance, the question "*How much time do you typically spend at home in each hour of the day?*", can be used to quantify the vectors $\boldsymbol{v}$ of Equation (4), needed to obtain P. The similarity of these vectors gives already a good estimation of the probability of interaction of the household members. Even if our experiment focused only on the household contacts, we envision that this approach can be directly extended to other settings, designing context-related contact matrices as done in [26]. Moreover, one can think of providing a finer estimation of $\boldsymbol{v}$ considering a multi-day average, so that $v_t \in [0, 1]$ is a probability to be at home (or, more generally, in a given place) at time t. The question "*How much of this time do you typically spend in isolation?*" then can allow one to re-weight the entries v_t to account for an actual probability of encounter. The last question "*How many face-to-face interactions you have per day?*" is an example of how one can quantify an individuals' activity rate. Given these estimates, the age parameters are obtained simply aggregating them according to the relevant age-group to obtain the activity vector $\boldsymbol{u}$.

4 Conclusion

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 combining behavioral parameters with demographic ones. While it comes as no surprise that a generalization of the matrix T would lead to better estimates, the most important aspects of our results are listed as follows:

- Simple, environment-independent models can accurately estimate HCM. The high quality and size of the PHIRST dataset gave us great 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
- Our proposed models are highly interpretable. We expect its parameters to be easily estimated with surveys, addressing questions such as those listed in Section 1. We expect this to be one of the significant outcomes of our research as we identified some practical questions to calibrate our model, bypassing proximity sensors.
- 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 aspect naturally reduces the number of parameters of the model, making the estimation process simpler and addresses the important requirement for surveys that the questions asked should have a simple answer.

The questions suggested in Section 1 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 main limitations of our methodology 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 waves of measurements in households were made asynchronously. After the cleaning procedure, it emerged that, as a consequence of the adoption of this choice for the scheduling of data collection in the field, weekdays and weekends are not evenly distributed among households and changes in the measured HCM 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 in the seasonality of the contact patterns, except for the duration of contact distribution shown in Figure 2c. 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, during a quarantine.

In conclusion, our study proposes a parametric model to estimate contact matrices with high accuracy. It improves over the purely demographic models in terms of accuracy and over the purely survey-based approaches in terms of simplicity of the data collection.

Given its simplicity and interpretability, we envision that our framework can be adopted to estimate contact matrices beyond the household setting. As a practical application, our results can impact the strategy to design the surveys currently adopted to quantify social contacts to mitigate the Covid19 and similar epidemics [47, 48].

Data availability

The contact matrices aggregated at the household level are made available at github.com/lorenzodallamico/PHIRST_CM.

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, from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101016233 (PERISCOPE) and from Fondation Botnar. LG, MT and LO 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.

References

1. Wallinga J, Edmunds WJ, Kretzschmar M. Perspective: human contact patterns and the spread of airborne infectious diseases. TRENDS in Microbiology. 1999;7(9):372–377.
2. Anderson RM, May RM. Infectious diseases of humans: dynamics and control. Oxford university press; 1992.
3. Meyers L. Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. Bulletin of the American Mathematical Society. 2007;44(1):63–86.
4. Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. The Lancet infectious diseases. 2020;20(6):669–677.
5. Walker PG, Whittaker C, Watson OJ, Baguelin M, Winskill P, Hamlet A, et al. The impact of COVID-19 and strategies for mitigation and suppression in low-and middle-income countries. Science. 2020;369(6502):413–422.
6. Sun K, Wang W, Gao L, Wang Y, Luo K, Ren L, et al. Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. Science. 2021;371(6526):eabe2424.
7. House T, Keeling M. Household structure and infectious disease transmission. Epidemiology & Infection. 2009;137(5):654–661.
8. Goeyvaerts N, Santermans E, Potter G, Torneri A, Van Kerckhove K, Willem L, et al. Household members do not contact each other at random: implications for infectious disease modelling. Proceedings of the Royal Society B. 2018;285(1893):20182201.
9. McCarthy Z, Xiao Y, Scarabel F, Tang B, Bragazzi NL, Nah K, et al. Quantifying the shift in social contact patterns in response to non-pharmaceutical interventions. Journal of Mathematics in Industry. 2020;10(1):1–25.

10. Cencetti G, Santin G, Longa A, Pigani E, Barrat A, Cattuto C, et al. Digital proximity tracing on empirical contact networks for pandemic control. Nature communications. 2021;12(1):1–12.
11. Wallinga J, Teunis P, Kretzschmar M. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. American journal of epidemiology. 2006;164(10):936–944.
12. Hilton J, Keeling MJ. Incorporating household structure and demography into models of endemic disease. Journal of the Royal Society Interface. 2019;16(157):20190317.
13. Li W, Zhang B, Lu J, Liu S, Chang Z, Peng C, et al. Characteristics of household transmission of COVID-19. Clinical Infectious Diseases. 2020;71(8):1943–1946.
14. Edmunds WJ, O'callaghan C, Nokes D. Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections. Proceedings of the Royal Society of London Series B: Biological Sciences. 1997;264(1384):949–957.
15. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS medicine. 2008;5(3):e74.
16. Prem K, Cook AR, Jit M. Projecting social contact matrices in 152 countries using contact surveys and demographic data. PLoS computational biology. 2017;13(9):e1005697.
17. Mistry D, Litvinova M, Pastore y Piontti A, Chinazzi M, Fumanelli L, Gomes MF, et al. Inferring high-resolution human mixing patterns for disease modeling. Nature communications. 2021;12(1):1–12.
18. Potter GE, Handcock MS, Longini Jr IM, Halloran ME. Estimating within-household contact networks from egocentric data. The annals of applied statistics. 2011;5(3):1816.
19. Smieszek T, Burri EU, Scherzinger R, Scholz RW. Collecting close-contact social mixing data with contact diaries: reporting errors and biases. Epidemiology & infection. 2012;140(4):744–752.
20. Mastrandrea R, Barrat A. How to estimate epidemic risk from incomplete contact diaries data? PLoS computational biology. 2016;12(6):e1005002.
21. Johnstone-Robertson SP, Mark D, Morrow C, Middelkoop K, Chiswell M, Aquino LD, et al. Social mixing patterns within a South African township community: implications for respiratory disease transmission and control. American journal of epidemiology. 2011;174(11):1246–1255.
22. Kiti MC, Kinyanjui TM, Koech DC, Munywoki PK, Medley GF, Nokes DJ. Quantifying age-related rates of social contact using diaries in a rural coastal population of Kenya. PloS one. 2014;9(8):e104786.
23. de Waroux OlP, Cohuet S, Ndazima D, Kucharski A, Juan-Giner A, Flasche S, 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. 2018;18(1):1–12.
24. Thindwa D, Jambo KC, Ojal J, MacPherson P, Phiri MD, Pinsent A, et al. Social mixing patterns relevant to infectious diseases spread by close contact in urban Blantyre, Malawi. Epidemics. 2022; p. 100590.
25. Naghavi M, Abajobir AA, Abbafati C, Abbas KM, Abd-Allah F, Abera SF, 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. 2017;390(10100):1151–1210.
26. Fumanelli L, Ajelli M, Manfredi P, Vespignani A, Merler S. Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread. 2012;.
27. Manna A, Dall'Amico L, Tizzoni M, Karsai M, Perra N. Generalized contact matrices for epidemic modeling; 2023.
28. Cattuto C, Van den Broeck W, Barrat A, Colizza V, Pinton JF, Vespignani A. Dynamics of person-to-person interactions from distributed RFID sensor networks. PloS one. 2010;5(7):e11596.
29. Stehlé J, Voirin N, Barrat A, Cattuto C, Isella L, Pinton JF, et al. High-resolution measurements of face-to-face contact patterns in a primary school. PloS one. 2011;6(8):e23176.
30. Vanhems P, Barrat A, Cattuto C, Pinton JF, Khanafer N, Régis C, et al. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. PloS one. 2013;8(9):e73970.
31. Ozella L, Gesualdo F, Tizzoni M, Rizzo C, Pandolfi E, Campagna I, et al. Close encounters between infants and household members measured through wearable proximity sensors. PloS one. 2018;13(6):e0198733.
32. Starnini M, Lepri B, Baronchelli A, Barrat A, Cattuto C, Pastor-Satorras R. Robust modeling of human contact networks across different scales and proximity-sensing techniques. In: International Conference on Social Informatics. Springer; 2017. p. 536–551.

33. Kiti MC, Melegaro A, Cattuto C, Nokes DJ. 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. 2019;4.

34. Cohen C, McMorrow ML, Martinson NA, Kahn K, Treurnicht FK, Moyes J, 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. 2021;15(6):789–803.

35. Kleynhans J, Tempia S, McMorrow ML, von Gottberg A, Martinson NA, Kahn K, 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. 2021;21(1):1–10.

36. Cohen C, Kleynhans J, von Gottberg A, McMorrow ML, Wolter N, Bhiman JN, 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;.

37. Kleynhans J, Tempia S, Wolter N, von Gottberg A, Bhiman JN, Buys A, 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. 2021;27(12):3020.

38. Cohen C, Kleynhans J, Moyes J, McMorrow ML, Treurnicht FK, Hellferscee O, 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. 2021;9(6):e863–e874.

39. Thindwa D, Wolter N, Pinsent A, Carrim M, Ojal J, Tempia S, 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. 2021;17(12):e1009680.

40. Wilkinson E, Giovanetti M, Tegally H, San JE, Lessells R, Cuadros D, et al. A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. Science. 2021;374(6566):423–431.

41. Igboh LS, Mcmorrow M, Tempia S, Emukule GO, Talla Nzussouo N, Mccarron M, et al. Influenza surveillance capacity improvements in Africa during 2011-2017. Influenza and other respiratory viruses. 2021;15(4):495–505.

42. Tempia S, Walaza S, Bhiman JN, McMorrow ML, Moyes J, Mkhencele T, 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. 2021;26(29):2001600.

43. Hoang T, Coletti P, Melegaro A, Wallinga J, Grijalva CG, Edmunds JW, et al. A systematic review of social contact surveys to inform transmission models of close-contact infections. Epidemiology (Cambridge, Mass). 2019;30(5):723.

44. Barabasi AL. The origin of bursts and heavy tails in human dynamics. Nature. 2005;435(7039):207–211.

45. Lance GN, Williams WT. Computer programs for hierarchical polythetic classification ("similarity analyses"). The Computer Journal. 1966;9(1):60–64.

46. Perra N, Gonçalves B, Pastor-Satorras R, Vespignani A. Activity driven modeling of time varying networks. Scientific reports. 2012;2(1):469.

47. Verelst F, Hermans L, Vercruysse S, Gimma A, Coletti P, Backer JA, 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. 2021;19(1):1–7.

48. Koppeschaar CE, Colizza V, Guerrisi C, Turbelin C, Duggan J, Edmunds WJ, et al. Influenzanet: citizens among 10 countries collaborating to monitor influenza in Europe. JMIR public health and surveillance. 2017;3(3):e7429.

S Supplementary information: Appendix

S.1 Data collection and pre-processing

Proximity data are measured with the `SocioPatterns` sensors that we here introduce, addressing the interested reader to [28] for a more detailed reference. Their functioning is based on the emission of low-power 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 related to the distance between the two sensors and can be used to filter suitably-defined close-range proximity relations. Additionally, each sensor periodically records some status properties that log metadata and diagnostic information. Among these, an accelerometer allows one to know every 15 minutes if the sensor is moving or not. Given the sensitivity of the accelerometer and the time-scale at which it operates, one can assume that if the sensor is still, then it is not worn. The cleaning procedure is summarized 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 inside a drawer

2. Contacts are filtered and only those with a suitable attenuation threshold. 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 beginning of the deployment (as reported in the diaries) 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.

4. The data collected by the sensors contains information on the hardware identification code. A mapping relates this identifier with the individuals' pseudonym that allows us to relate contacts and metadata. 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 sufficiently 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 great excess of activity during night hours was observed in three households (yellow dots of Figure S.1a) during the first deployment. This may occur, for instance, if the sensors are left in proximity on a vibrating surface: the accelerometer filter does not remove these contacts even though the sensors were not worn at that moment.

Only the households in which all three deployments led to valid measurements (all green dots in Figure S.1a) were included in our study. Figure S.1b, c, d, e further show

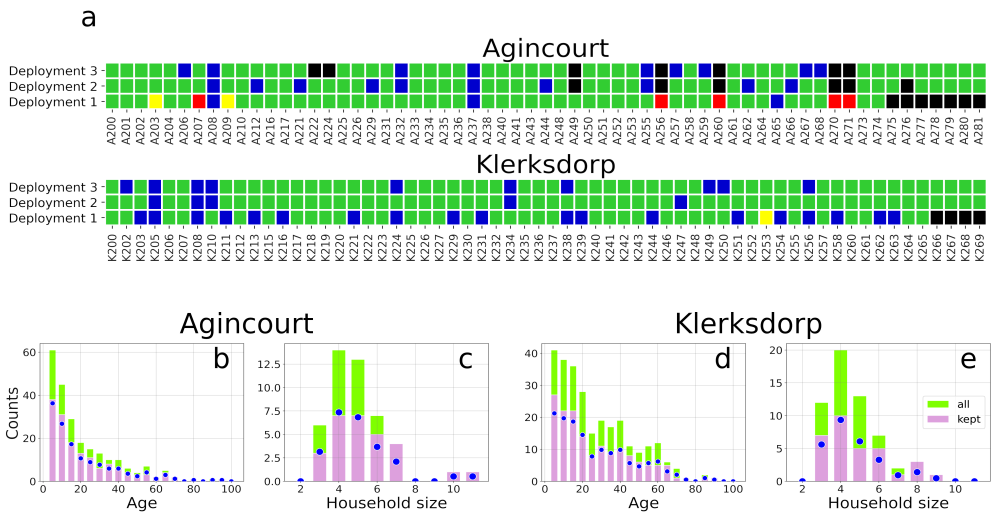

Fig S.1. Raw data characteristics. **a**: data quality. On the x-axis we plot households, while on the y-axis the deployments. For each (household-deployment) we assign a color code: black indicates that the household did not participate; red that all household's sensors had data quality issues and did 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 refers to the 60 households with valid measurements in all three deployments (see **a**). Blue dots are obtained by multiplying the height of the green bars for the fraction of the included households, that is the expected bar height, given the cleaned dataset size. **c** and **e**: household size distribution. Legends and colors follow **b** and **d**.

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.

S.2 Validation approach

Sampling the villages

First of all, in order to devise a good approximation of HCM, it is necessary to define a suitable distance to compare them. 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, $\Phi_a = 0$. In some extreme cases there is no way to consistently compare HCM 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 groups of HCM (*villages*) $\mathcal{X}$, *i.e.* small groups of household-deployment (h, d) pairs that guarantee that $\Phi_a > 0$ for all a. To build the samples $\mathcal{X}$ we then first select some (h, d) at random, with the constraint to achieve $\Phi_a > 0$ 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 empirically choose $|\mathcal{X}| = 8$ because it is 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 $\Phi_a > 0$ 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.

Model calibration

Given the samples of villages we now compute the value $\boldsymbol{u}$ as the result of the following optimization problem

$$\boldsymbol{u} = \underset{\boldsymbol{v}\ :\ \boldsymbol{v}^T\mathbf{1}=\text{const}}{\arg\min} \quad d_{\mathrm{C}}\left(C, T \circ \boldsymbol{v}\boldsymbol{v}^T\right),$$

where $\left[T \circ (\boldsymbol{v}\boldsymbol{v}^T)\right]_{ab} = T_{ab}v_a v_b$ and d_{C} is a modified Canberra distance. Let A, B be two symmetric matrices of size n_{age}, then

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

where $\tilde{A}$ is the matrix A divided by its mean (and equivalently $\tilde{B}$ is B divided by its mean). The distance d_{C} is the Canberra distance computed on the matrices $\tilde{A}, \tilde{B}$, instead of A, B, hence we refer to it as *modified Canberra distance*. This choice of the distance is motivated by the two following points

1. The entries of C may differ even by a factor 100 as shown in Figure 4. 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. For this reason it is unsuited for an optimization as it would poorly estimate the small entries of C. On the opposite, the relative distance d_{C} gives approximately the same weight to all matrix entries and can be used for this purpose.

2. The modified Canberra distance compares a normalized version of the contact matrices because we are interested in determining them up to a constant factor. We then have for any $\alpha, \beta > 0$, $d_{\mathrm{C}}(\alpha A, \beta B) = d_{\mathrm{C}}(A, B)$.

Occupation parameter

We here detail the strategy to determine the vectors $\boldsymbol{y}, \boldsymbol{\eta}$ appearing in Figure 3b, 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 is the average of the Boolean indicator for all people of age a in $\mathcal{X}$.