Do contacts over distance follow a power-law distribution? Estimation of the social contact distance kernel

Kim Van Kerckhove[1], Christel Faes[1], Philippe Beutels[23], Niel Hens[12]

[1] I-Biostat, Hasselt University, Belgium
[2] CHERMID, VAXINFECTIO University of Antwerp, Belgium
[3] School of Public Health and Community Medicine, The University of New South Wales, Australia

E-mail for correspondence: kim.vankerckhove@uhasselt.be

Abstract: Do contacts over distance show a multimodal form, with a peak of contacts close to home and a second peak further away from home, or is a power-law form sufficient? By using data from our social contact study, we were able to test this hypothesis. We exploited various distributions for the contacts at a certain distance, e.g. Poisson, Negative Binomial, ..., and incorporated random effects to account for the clustering of contacts within participants. Various forms of the underlying distribution were tested by integrating their information into the observed categories. The preliminary results support a Weibull form for the distribution of contacts over distance, however subtle differences are present when differentiating by the participant's age and by week or weekend days.

Keywords: Contacts; Distance kernel; Power-law; Random effects.

1 Introduction

Cooper (2006) noticed the lack of appropriately incorporating spatial information about contacts in epidemic models. He reported a study by Riley and Ferguson (2006) in which journey-to-work data were used as a proxy for the spatial distribution of potential contacts, a rough approximation to the truth according to Cooper (2006). He furthermore referred to unpublished work on contacts over distance to indicate the possible biases in using commuting data. Read et al. (2014) noted the same gap and presented data on contacts over distance in the setting of southern China. Their results

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

consisted of comparing the contacts over distance by rural and urban areas, describing the differences by age, but nonetheless no estimation of the distance kernel itself. Differences in the number of contacts over distance by age were shown, whereby the age group between 20 and 29 years of age was most mobile. The estimation of the distance kernel could lead to more realistic models and hence would possibly fill in the gap noted by previous authors. The current study aims at estimating the underlying distribution of contacts over distance, hence the distance kernel, while accounting for clustering and allowing differentiations based on characteristics of the participants.

2 Methodology

Using the time use data and the contact data from our social contact survey (described in detail by Willem et al., 2012) a probabilistic match between contacts and their distances from home was obtained. In the time use part, the participants indicated the place and at which distance they spent most of their time for certain time blocks (e.g. 5-8 h, 8-9 h, 9-10 h, ...). The distance (from home) was divided into 4 categories: 0-1 km, 2-9 km, 10-74 km and more than 75 km. The contact diary part was similar to previous contact surveys (e.g. Mossong et al., 2008), hence information on the location of the contact was available. A contact was defined as either a conversation with a person at less than 3 meters distance or a person they touched (skin-to-skin contact), hence contacts by phone or internet were not recorded. Repeated contacts were collected only once by combining the information. To combine the two parts, contacts were separated by location, if multiple locations were reported. The information from the time use data was summarized to have consecutive time blocks that either differed in the location the participant spent most of their time or by the distance at which the participant spent their time. Contacts at specific locations were further separated to have each of the possible options present from the time use data (by time and distance). As contacts might occur in a short period of time and the time spent at this location might not be of substantial duration, it occurs that a contact cannot be linked to a distance nor time. This is a drawback of the study design, necessitating the omission of these data. Since the probabilistic matching involves uncertainty, weights are given to each link. We refer to a combination of possible links as a distributed contact, and the weights sum to one for each distributed contact.
The information present for a possible link i can be described by $\mathbf{D_i} = (D_{1i}, D_{2i}, D_{3i}, D_{4i})$ with D_{1i} equal to 1 if the link occurs at distance 0-2 km and 0 otherwise. Each element of this stochastic vector follows a Poisson (Negative Binomial) distribution. Taking the weights for each link

into account, the following log-likelihood will be used:

$$\mathrm{ll}(\boldsymbol{\beta}|\mathbf{d}) = \sum_{i=1}^{p}\sum_{j=1}^{4}\left[-\lambda_{ji} + y_{ij}\log(\lambda_{ji}) - \log(y_{ij}!)\right].$$

Information concerning the participant (age, information recorded on week or weekend day) can be incorporated in the model via the parameters and the link functions. Let η_{1i}, η_{2i}, η_{3i}, and η_{4i} denote the linear predictors, which include the information of the participant through $\boldsymbol{\beta}$. These β-estimates can differ for each linear predictor. The parameters can be obtained as follows (with $n_i = \sum_{j=1}^{k} w_{ij}$).

$$\begin{aligned} \lambda_{1i} &= n_i \exp(\eta_{1i}), & \lambda_{2i} &= n_i \exp(\eta_{2i}) \\ \lambda_{3i} &= n_i \exp(\eta_{3i}), & \lambda_{4i} &= n_i \exp(\eta_{4i}). \end{aligned}$$

However, this approach does not take the clustering of contacts within participants into account. Random effects were added for each linear predictor. Furthermore, one can assume an underlying distribution for the contacts over distance, say $f(u, \boldsymbol{\theta})$. The linear predictors include the expression of the underlying distribution as for example $\eta_2 = \int_2^{10} f(u, \boldsymbol{\theta})\mathrm{d}u$. Information concerning the participant are incorporated in the parameters $\boldsymbol{\theta}$ of the underlying distribution. Various distributions including Powerlaw and Weibull are considered for $f(u, \boldsymbol{\theta})$.
Additionally, various models for the total number of contacts were compared to obtain the best fit. As such we are able to combine these two elements in the estimation of the joint density of the total number of contacts and the number of contacts over distance. This density allows us to test the assumption of independence between the total number of contacts and contacts over distance.

3 Preliminary results

A total of 41,327 links were recorded by 1527 participants, after removing the links with missing values for distance. Taking the weighted sum of the distributed contacts shows that most of the contacts (10,119.17) were reported at distance 2-9 km from home, followed by contacts close to home (8374.84) and contacts 10-74 km from home (7567.59). Substantially fewer contacts were reported at a distance of more than 75 km from home (837.51).
In a first attempt, we model the parameters allowing different estimates for various age classes and week or weekend days, but ignoring the clustering within participants. The age classes considered are based on the schooling system: $[0, 3)$, $[3, 6)$, $[6, 12)$, $[12, 18)$, $[18, 25)$, $[25, 45)$, $[45, 65)$ and $[65, 100)$. The saturated model, which allows a different estimate for each age group and week/weekend, has the lowest AIC (-65150) when age and week were

included together with an interaction term. Using an underlying Weibull model lead to the best fit from the various underlying distributions (AIC=-64081), however the saturated model outperforms this model. A further improvement in the Weibull model is to allow the two parameters to vary by age and week.
The estimates for the probabilities ($E(\frac{Y_1}{n_i})$) based on the saturated model showed a discrepancy over age: children showed a large proportion (around 55%) of contacts at home, whereas adults or young adults showed a smaller proportion of contacts close to home (around 25-35%), but an increase in the proportion of contacts at distance 10-74 km from home (around 29-40% for young adults and adults compared to 9-14% for children). Furthermore, during weekends generally a larger proportion of the contacts were reported closer to home (mostly around 30-59%).

4 Discussion

The preliminary results indicate a Weibull underlying distribution, however an extension of this model should give more evidence but clear influences of the participant's age and the day of collection are present. These subtle differences should be incorporated in future models such as agent based models or meta-population models in infectious diseases. Due to the study design not all contacts were linked to a possible distance and hence information was lost.

References

Cooper, B. (2006). Poxy models and rash decisions. *Proceedings of the National Academy of Sciences of the US*, **103**, 12221-12222.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., et al. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine*, **5(3)**, e74.

Read, J.M., Lessler, J., Riley, S., Wang, S., Tan, L.J., Kwok, K.O., Guan, Y., Jiang, C.Q., and Cummings, D.A.T. (2014). Social mixing patterns in rural and urban areas of southern China. *Proceedings of the Royal Society B*, **281**, 20140268.

Riley, S. and Ferguson, N.M. (2006). Smallpox transmission and control: Spatial dynamics in Great Britain. *Proceedings of the National Academy of Sciences of the US*, **103**, 12637-12642.

Willem, L., Van Kerckhove, K., Chao, D.L., Hens, N., and Beutels, P. (2012). A nice day for an infection? Weather conditions and social contact patterns relevant to influenza transmission. *PLoS ONE*, **7**, e48695.