The Journal of Open Source Software

conmat: generate synthetic contact matrices for a given age-stratified population

Nicholas Tierney[1¶], **Chitra Saraswati**[1], **Aarathy Babu**[1], **Michael Lydeamore**[3], **and Nick Golding**[1,2]

1 The Kids Research Institute Australia, WA, Australia **2** University of Western Australia, WA, Australia **3** Department of Econometrics and Business Statistics, Monash University, VIC, Australia ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Charlotte Soneson

Reviewers:

- @anthbapt
- @Bisaloo

Submitted: 10 March 2025
Published: unpublished

Summary

Contact matrices describe the number of contacts between individuals. They are used to create models of infectious disease spread. conmat is an R package which generates synthetic contact matrices for arbitrary input demography, ready for use in infectious disease modelling.

There are currently few options for a user to access synthetic contact matrices (Funk et al., 2018; Prem et al., 2017). Existing code to generate synthetic contact matrices from Prem et al. (2017) are not designed for replicability, are restricted to select countries, and provide no sub-national demographic estimates.

The conmat package exposes model fitting and prediction separately to the user. Users can fit a model based on a contact survey such as POLYMOD (Mossong et al., 2008), then predict from this model to their own demographic data. This means users can generate synthetic contact matrices for any region, with any contact survey.

We demonstrate a use-case for conmat by creating contact matrices for sub-national level (in this case, a state) in Australia.

For users who do not wish to run the entire conmat pipeline, we have pre-generated synthetic contact matrices for 200 countries, based on a list of countries from the United Nations, using a model fit to the POLYMOD contact survey. These resulting synthetic contact matrices, and the associated code, can be found in the syncomat analysis pipeline (GitHub, Zenodo) (Saraswati et al., 2024).

Statement of need

Infectious diseases like influenza and COVID-19 spread via social contact. If we can understand patterns of contact—which individuals are more likely be in contact with each other—then we will be able to create models of how disease spreads. Epidemiologists and public policy makers can use these models to make decisions to keep a population safe and healthy.

Empirical estimates of social contact are provided by social contact surveys. These provide samples of the frequency and type of social contact across different settings (home, work, school, other).

A prominent contact survey is the POLYMOD study by Mossong et al. (2008), which surveyed 8 European countries: Belgium, Germany, Finland, Great Britain, Italy, Luxembourg, The Netherlands, and Poland (Mossong et al., 2008).

These social contact surveys can be projected on to a given demographic structure to produce estimated daily contact rates between age groups. These are known as ‘synthetic’ contact matrices. A widely used approach by Prem et al. (2017, 2021) produced synthetic contact matrices for 177 countries at ‘urban’ and ‘rural’ levels for each country.

However, there were major limitations with the methods in Prem et al. (2021). First, not all countries were included in their analyses. Second, the contact matrices only covered broad population groups within entire countries. This presents challenges for decision makers who are often working at a sub-national geographical scale, with differing demographic structure in different sub-populations. Third, the code provided by Prem et al. was not designed for replicability and easy modification with user-defined inputs.

The conmat package was developed to fill the specific need of creating contact matrices for arbitrary age categories and populations (as shown in the below example) to inform infectious diease models. We developed the method primarily to output synthetic contact matrices. We also provided methods to create next generation matrices for modelling.

Example

We will generate a contact matrix for Tasmania, a state in Australia, using a model fitted from the POLYMOD contact survey. We can get the age-stratified population data for Tasmania from the Australian Bureau of Statistics (ABS) with the helper function, `abs_age_state()`:

```
tasmania <- abs_age_state("TAS")
head(tasmania)
```

```
# A tibble: 6 × 4 (conmat_population)
 - age: lower.age.limit
 - population: population
   year state lower.age.limit population
  <dbl> <chr>           <dbl>      <dbl>
1  2020 TAS                 0      29267
2  2020 TAS                 5      31717
3  2020 TAS                10      33318
4  2020 TAS                15      31019
5  2020 TAS                20      31641
6  2020 TAS                25      34115
```

We can then generate a synthetic contact matrix for Tasmania, by extrapolating the contact patterns between age groups learned from the POLYMOD study, using `extrapolate_polymod()`.

```
tasmania_contact <- extrapolate_polymod(population = tasmania)
tasmania_contact
```

We can plot the resulting contact matrix for Tasmania with `autoplot`, shown in (Figure 1).

```
autoplot(tasmania_contact)
```

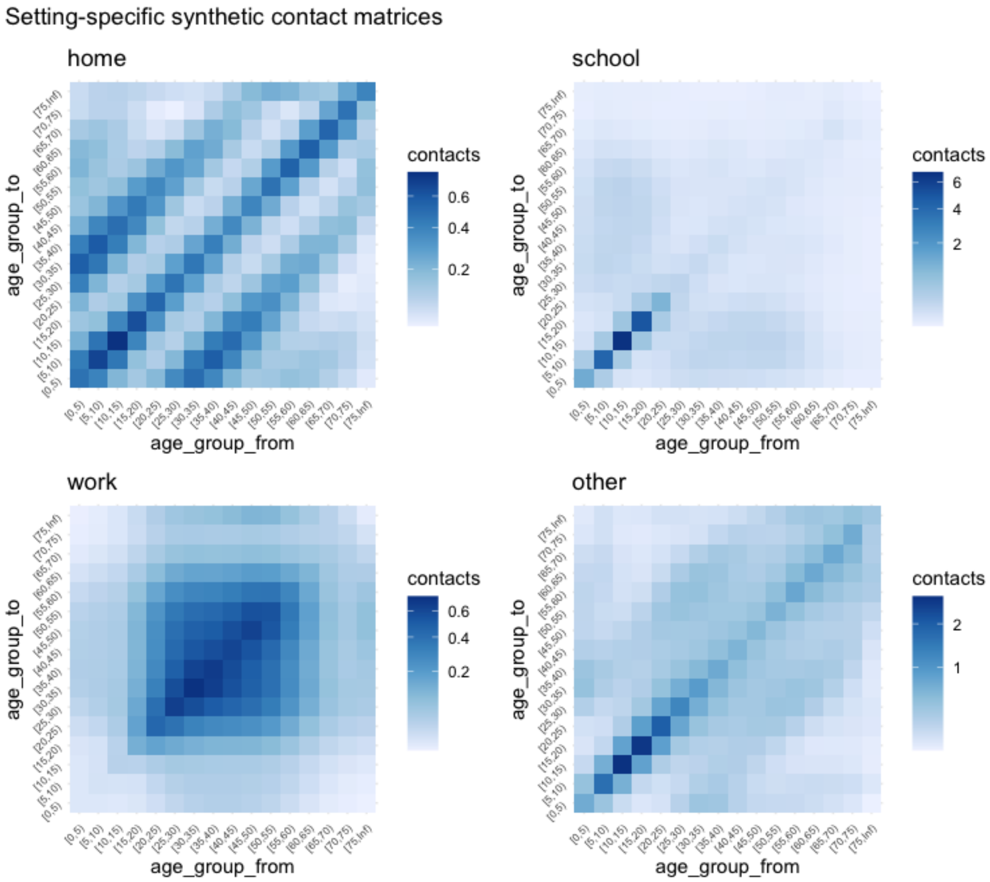

Figure 1: Contact patterns between individuals for different age groups across four settings: home, work, school, and other. The x axis shows the age groups for focal individuals ('from'), and the y axis shows the age groups for people those individuals have contact with ('to'), coloured by the average number of contacts the individual has in that age group. We see different contact patterns in different settings, for example, diagonal with 'wings' for the home setting.

Implementation

The overall approach of `conmat` has two parts:

1) fit a model to predict individual contact rates, using an existing contact survey;
2) predict a synthetic contact matrix using age population data.

Model fitting

`conmat` was built to predict at four settings: work, school, home, and other. One model is fitted for each setting. Each model fitted is a Poisson generalised additive model (GAM) with a log link function, which predicts the count of contacts, with an offset for the log of participants. The model has six covariates to explain six key features of the relationship between ages, and two optional covariates for attendance at school or work. The two optional covariates are included depending on which setting the model is fitted for.

Each cell in the resulting contact matrix (after back-transformation of the link function), indexed $(i,\ j)$, is the predicted number of people in age group j that a single individual in age group i will have contact with per day. The sum over all age groups j for a particular age group i is the predicted total number of contacts per day for each individual of age group i.

The six covariates are:

- $|i-j|$,
- $|i-j|^2$,
- $i+j$,
- $i \times j$,
- $\max(i,j)$ and
- $\min(i,j)$.

These covariates capture typical features of inter-person contact, where individuals primarily interact with people of similar age (the diagonals of the matrix), and with grandparents and/or children (the so-called 'wings' of the matrix). The key features of the relationship between the age groups, represented by spline transformations of the six covariates, are displayed in Figure 2 for the home setting. The spline-transformed $|i-j|$ and $|i-j|^2$ terms give the strong diagonal lines modelling people generally living with those of similar age and the intergenerational effect of parents and (faintly) grandparents with children. The spline-transformed $\max(i,j)$ and $\min(i,j)$ terms give the higher rates of at-home contact among younger people of similar ages and with their parents.

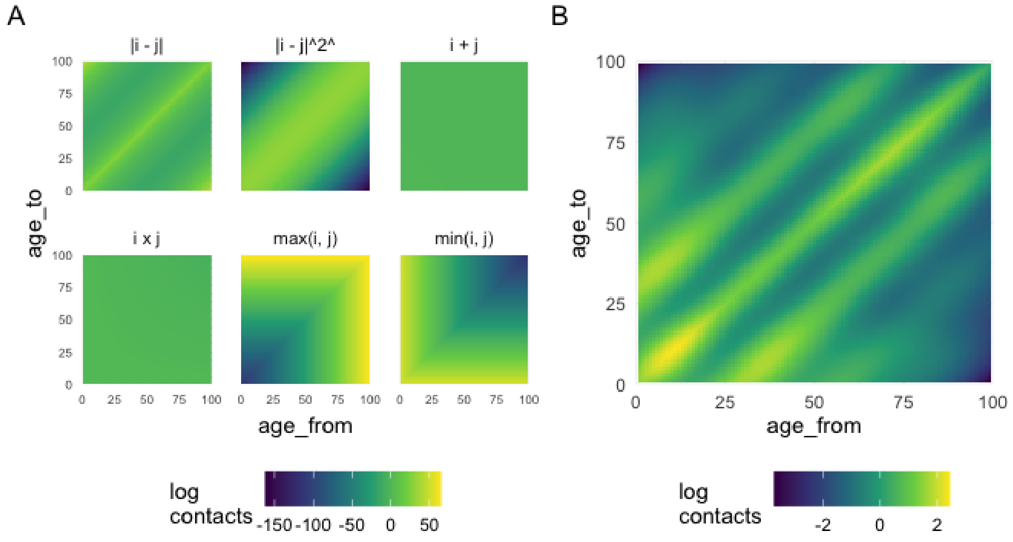

Figure 2: Partial predictive plot (A) and overall synthetic contact matrix (B) for the Poisson GAM fitted to the POLYMOD contact survey in the home setting. The strong diagonal elements, and parents/grandparents interacting with children result in the classic 'diagonal with wings' shape.

Visualising the partial predictive plots for other settings (school, work and other) show patterns that correspond with real-life situations. A full visualisation pipeline is available at https://idem-lab.github.io/conmat/dev/articles/visualising-conmat.html

Conclusions and future directions

The conmat software provides a flexible interface to generating synthetic contact matrices using population data and contact surveys. These contact matrices can then be used in infectious disease modelling and surveillance.

The main strength of conmat is its interface requiring only age population data to create a synthetic contact matrix. Current approaches provide only a selection of country level contact matrices. This software can predict to arbitrary demography, such as sub-national or simulated populations.

We provide a trained model of contact rate that is fit to the POLYMOD survey for ease of use. The software also has an interface to train models to other contact surveys, such as Jarvis et al. (2024). This is important as POLYMOD represents contact patterns in 8 countries in Europe, and contact patterns are known to differ across nations and cultures.

The covariates used by conmat were designed to represent the key features that are typically present in a contact matrix for different settings (work, school, home, other). Including other sources of information that may better describe these contact patterns, such as inter-generational mixing, or differences in school ages of a local demographic, may improve model performance.

The interface to the model formula in conmat is fixed; users cannot change the covariates of the model. This means if there is an unusual structure in their contact data it might not be accurately captured by conmat. This fixed formula was a design decision made to focus on the key feature of conmat: using only age population data to predict a contact matrix.

Public health decisions are often based on age specific information, which means the more accurate your age specific models are, the better those decisions are likely to be. This is the first piece of software that will provide appropriate contact matrices for a population, which means more accurate models of disease.

This code underlying this software was used as a key input into several models for COVID-19 transmission and control in Australia and contributed to decisions around vaccination policy (Conway et al., 2023; McVernon et al., n.d.; Ryan et al., 2024).

Some future directions for this software include:

- integrating model fitting methods with contact survey prepared using the socialmixr R package (Funk et al., 2018);
- enabling prediction to arbitrary age brackets, e.g. monthly age bins for infants;
- fitting to multiple contact surveys simultaneously, e.g., POLYMOD and CoMix;
- providing estimates of uncertainty in contact matrices;
- adding methods to include household size distributions in predictions of contacts in the 'home' setting;

Software is never finished, and the software in its current format has proven useful for infectious disease modelling. In time we hope it can become more widely used and be useful for more applications in epidemiology and public health.

References

Conway, E., Walker, C. R., Baker, C., Lydeamore, M. J., Ryan, G. E., Campbell, T., Miller, J. C., Rebuli, N., Yeung, M., Kabashima, G., Geard, N., Wood, J., McCaw, J. M., McVernon, J., Golding, N., Price, D. J., & Shearer, F. M. (2023). COVID-19 vaccine coverage targets to inform reopening plans in a low incidence setting. *Proceedings of the Royal Society B: Biological Sciences*, *290*(2005), 20231437. https://doi.org/10.1098/rspb.2023.1437

Funk, S., Willem, L., & Gruson, H. (2018). *Socialmixr: Social mixing matrices for infectious disease modelling*. https://doi.org/10.32614/CRAN.package.socialmixr

Jarvis, C., Coletti, P., Backer, J., Munday, J., Faes, C., Beutels, P., Althaus, C., Low, N., Wallinga, J., Hens, N., & Edmunds, J. (2024). *CoMix data (last round in BE, CH, NL and UK)* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.11154066

McVernon, J., McCaw, J., Tierney, N., Miller, J., Lydeamore, M., Golding, N., Shearer, F., Geard, N., Cameron, Z., Baker, C., Walker, C., Ross, J., Wood, J., Conway, E., & Mueller, I. (n.d.). *Doherty institute - modelling*. Retrieved August 19, 2024, from https://www.doherty.edu.au/our-work/institute-themes/viral-infectious-diseases/covid-19/covid-19-modelling/modelling

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., & Edmunds, W. J. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLOS Medicine*, *5*(3), e74. https://doi.org/10.1371/journal.pmed.0050074

Prem, K., Cook, A. R., & Jit, M. (2017). Projecting social contact matrices in 152 countries using contact surveys and demographic data. *PLOS Computational Biology*, *13*(9), e1005697. https://doi.org/10.1371/journal.pcbi.1005697

Prem, K., Zandvoort, K. van, Klepac, P., Eggo, R. M., Davies, N. G., Group, C. for the M. M. of I. D. C. W., Cook, A. R., & Jit, M. (2021). Projecting contact matrices in 177 geographical regions: An update and comparison with empirical data for the COVID-19 era. *PLOS Computational Biology*, *17*(7), e1009098. https://doi.org/10.1371/journal.pcbi.1009098

Ryan, G. E., Shearer, F. M., McCaw, J. M., McVernon, J., & Golding, N. (2024). Estimating measures to reduce the transmission of SARS-CoV-2 in Australia to guide a "National Plan" to reopening. *Epidemics*, *47*, 100763. https://doi.org/10.1016/j.epidem.2024.100763

Saraswati, C. M., Lydeamore, M., Golding, N., Babu, A., & Tierney, N. (2024). *syncomat: Synthetic Contact Matrices for 200 UN Countries* (Version v1.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.11365943