The Journal of Open Source Software

ungroup: An R package for efficient estimation of smooth distributions from coarsely binned data

Marius D. Pascariu[1], **Maciej J. Dańko**[3], **Jonas Schöley**[1], **and Silvia Rizzi**[2]

1 Institute of Public Health, Center on Population Dynamics, University of Southern Denmark, Odense, Denmark **2** Institute of Public Health, Unit of Epidemiology Biostatistics and Biodemography, University of Southern Denmark, Odense, Denmark **3** Max Planck Institute for Demographic Research, Rostock, Germany

DOI: 10.21105/joss.00936

Software

- Review
- Repository
- Archive

Submitted: 10 September 2018
Published: 11 September 2018

`ungroup` is an open source software library written in the `R` programming language (R Core Team, 2018) that introduces a versatile method for ungrouping histograms (binned count data) assuming that counts are Poisson distributed and that the underlying sequence over a fine grid to be estimated is smooth. The method is based on the composite link model (Thompson & Baker, 1981) and estimation is achieved by maximizing a penalized likelihood (P. H. Eilers, 2007), which extends standard generalized linear models. The penalized composite link model (PCLM) implements the idea that observed counts, interpreted as realizations from Poisson distributions, are indirect observations of a finer (ungrouped) but latent sequence. This latent sequence represents the distribution of expected means on a fine resolution and has to be estimated from the aggregated data. Estimates are obtained by maximizing a penalized likelihood. This maximization is performed efficiently by a version of the iteratively re-weighted least-squares algorithm. Optimal values of the smoothing parameter are chosen by minimizing Bayesian or Akaike's Information Criterion (Hastie & Tibshirani, 1990).

Ungrouping binned data can be desirable for many reasons: Bins can be too coarse to allow for accurate analysis; comparisons can be hindered when different grouping approaches are used in different histograms; the last interval may be wide and open-ended masking the tail behaviour of the underlying distribution. Age-at-death distributions grouped into age classes and abridged life tables are examples of binned data which can be ungrouped with the package `ungroup`. The modest assumptions of the methodology underpinning the PCLM method make it suitable for many demographic and epidemiological applications. For a detailed description of the method and applications see Rizzi, Gampe, & Eilers (2015) and Rizzi et al. (2016).

The penalized composite link model can be extended to a two-dimensional regression problem (Rizzi et al., Forthcoming 2018). The two-dimensional regression analysis combines two approaches: the PCLM for ungrouping in one dimension and two-dimensional smoothing with P-splines (Currie, Durban, & Eilers, 2004). As an example one can ungroup age-specific distributions from the coarsely grouped data and smooth across adjacent calendar years to estimate both detailed age-at-death distributions and mortality time trends.

Acknowledgment

We thank Paul H.C. Eilers who provided insight and expertise that greatly supported the creation of this R package; and Catalina Torres and Tim Riffe for testing and offering feedback on the early versions of the software.

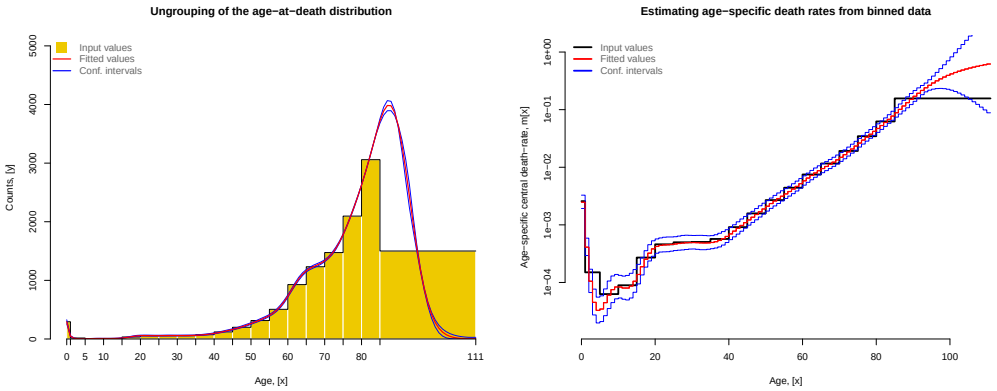

Figure 1: Ungrouping of the age-at-death distribution and estimating age-specific death rates. The original death counts and exposures taken from the Human Mortality Database (2018) using the `MortalityLaws` R package (2018) were grouped in 5-year bins plus a wide class for ages 85+. In each panel, the original aggregated data is compared with smoothly estimated values.

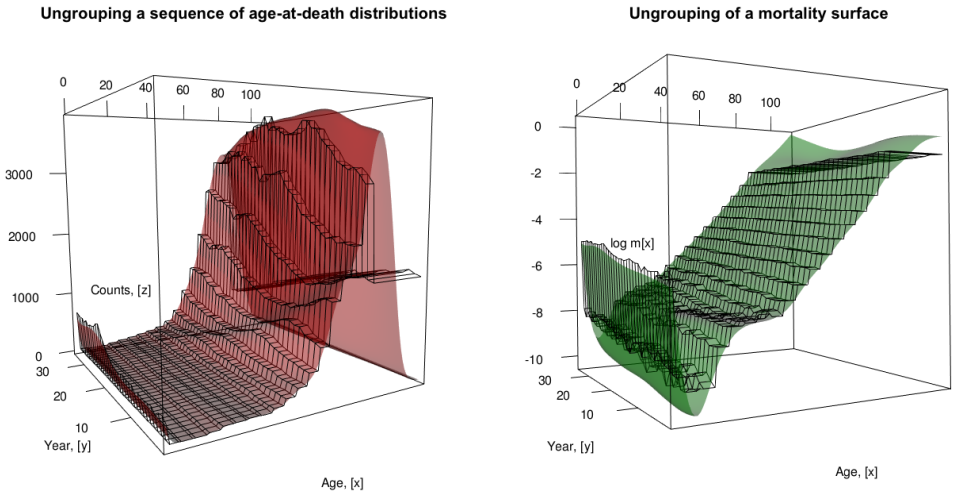

Figure 2: Two-dimensional ungrouping of the age-at-death distributions and mortality surface. The 3-D figures are generate using the `rgl` R package (2018).

The authors are grateful to the following institutions for their support:

- University of Southern Denmark;
- Max Planck Institute for Demographic Research;
- SCOR Corporate Foundation for Science.

References

Adler, D., Murdoch, D., & others. (2018). *Rgl: 3D visualization using opengl*. Retrieved from https://CRAN.R-project.org/package=rgl

Currie, I. D., Durban, M., & Eilers, P. H. (2004). Smoothing and forecasting mortality rates. *Statistical modelling*, *4*(4), 279–298. doi:10.1191/1471082X04st080oa

Eilers, P. H. (2007). Ill-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, *7*(3), 239–254. doi:10.1177/1471082X0700700302

Hastie, T. J., & Tibshirani, R. J. (1990). Generalized additive models. *Monographs on Statistics and Applied Probability*, *43*.

Human Mortality Database. (2018). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Data downloaded on 17/01/2018. Retrieved from https://www.mortality.org

Pascariu, M. D. (2018). *MortalityLaws: Parametric Mortality Models, Life Tables and HMD*. Retrieved from https://github.com/mpascariu/MortalityLaws

R Core Team. (2018). R: A language and environment for statistical computing [Internet]. Vienna, Austria. R version 3.5.0 (2018-04-23). Retrieved from https://www.r-project.org

Rizzi, S., Gampe, J., & Eilers, P. H. C. (2015). Efficient estimation of smooth distributions from coarsely grouped data. *American Journal of Epidemiology*, *182*(2), 138–147. doi:10.1093/aje/kwv020

Rizzi, S., Halekoh, U., Thinggaard, M., Engholm, G., Christensen, N., Johannesen, T. B., & Lindahl-Jacobsen, R. (Forthcoming 2018). How to estimate mortality trends from grouped vital statistics. *International Journal of Epidemiology*.

Rizzi, S., Thinggaard, M., Engholm, G., Christensen, N., Johannesen, T. B., Vaupel, J. W., & Lindahl-Jacobsen, R. (2016). Comparison of non-parametric methods for ungrouping coarsely aggregated data. *BMC medical research methodology*, *16*(1), 59. doi:10.1186/s12874-016-0157-8

Thompson, R., & Baker, R. (1981). Composite link functions in generalized linear models. *Applied Statistics*, 125–131. doi:10.2307/2346381