

Thesis proposal: A blended distance to define “people-like-me”

Anaïs Fopma (6199356)

Supervisors: Prof. dr. Stef van Buuren, Dr. Gerko Vink & Mingyang Cai

Programme: Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Word count: 723

Date: 15 October 2021

1. Introduction

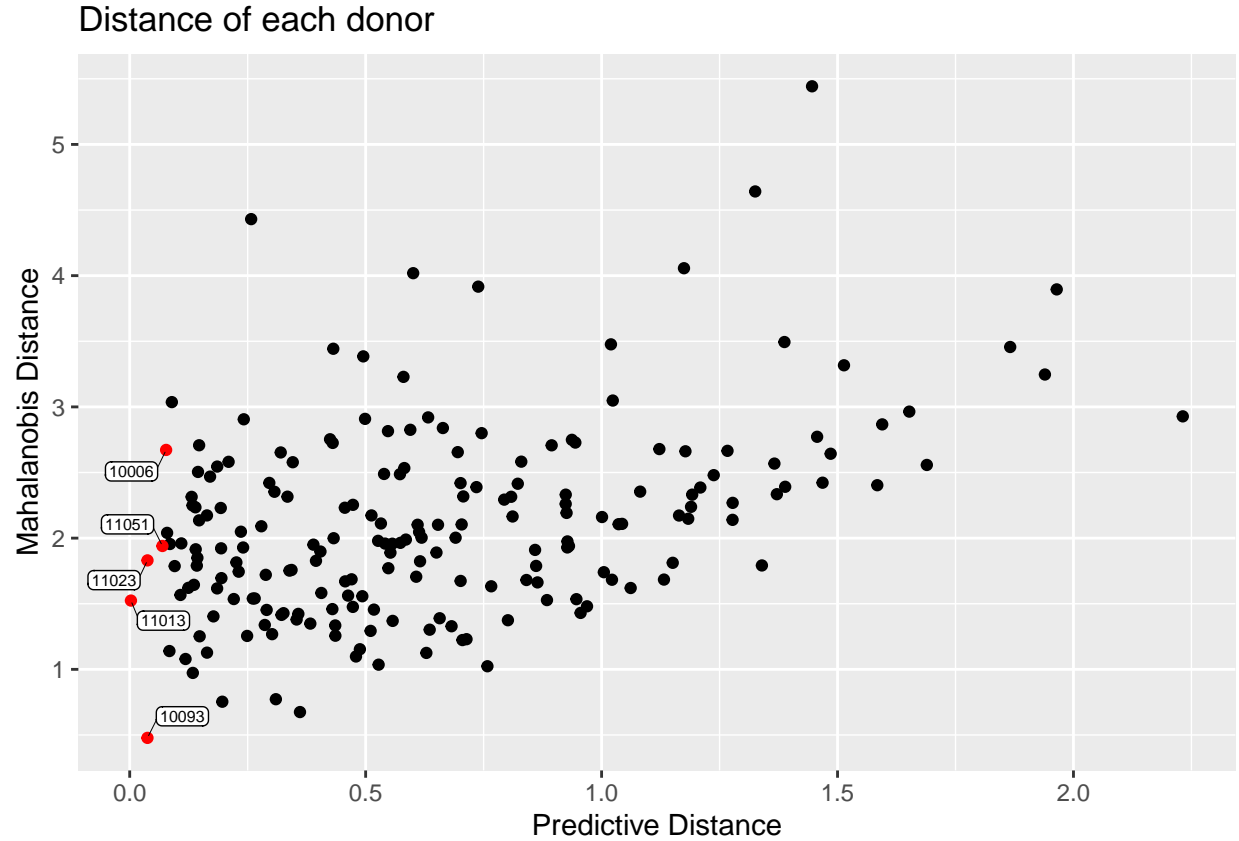
Growth curve modeling is a technique used to predict the future development of a child. It has important practical implementations, as it can provide answers to questions health professionals, parents, and insurance companies may have.¹ These are questions such as: ‘Given what is known about the child, how will it develop in the future?’ and ‘In the case of inhibited growth, will development be normal after intervention?’¹ Curve matching¹ is a growth curve modeling approach that is currently used for this purpose. The aim of curve matching is to predict the growth of a target child by using the data of older children of which data at a later age are already available in a donor database. To do this as accurately as possible, we use the data of a number of donors that are most similar to the target. These donors are the so-called “people like me”. In order to do this, we first need to define similarity and match the donors to the target. The key question is: How do we obtain good matches? The current approach uses predictive mean matching (PMM). First, a linear regression model is fitted on the donor database. Then, this model is used to predict the values for all donors in the database and for the target at a certain point in the future, for example at 14 months. Finally, the distance between the predicted value of each of the donors and the predicted value of the target is calculated. This is the predictive distance. A number of donors with the smallest predictive distance are selected as the best matches. Their growth curves are then plotted and point estimates can be calculated by averaging the measurements.

PMM is known to be extremely robust to violations of linearity and normality assumptions, very fast and the state of the art in missing data imputation.¹ However, there are two reasons to move beyond the predictive distance used in PMM and investigate alternative distance metrics. First, users of curve matching may find it difficult to select one particular future time point to base the matches on. Second, the predictive distance may make the matches look unconvincing. Since different profiles may lead to the same predicted value, the curves of some matches may be quite far from the curve of the target.

For these reasons, the practical use of curve matching can be improved by combining the predictive distance with alternative distance metrics that take into account historic similarity, thus creating a “blended distance” metric. It is expected that this may come at the cost of prediction accuracy, however. Therefore, the objective of this study is to investigate where to strike a balance between the predictive distance and alternative distances.

2. Strategy

A simulation study will be conducted to investigate what the properties of a blended distance measure would be. Different possibilities will be explored, each consisting of a combination of the predictive distance with an alternative distance metric, such as the Mahalanobis distance, Fréchet distance, and Hamming distance. To illustrate a possible combination, an example is provided in the figure below. We have the growth data of 200 children from the SMOCC study.² The first subject is taken as the target, the 199 other subjects as the donors. In the figure, the Mahalanobis distance of each donor for the measurements during the first six months of growth is plotted against the predictive distance between each donor and the target. The five matches based on the predictive distance³ are shown in red and labeled with the corresponding subject id’s. We can see that although these matches have a small predictive distance, some matches (especially subject 10006) have a large Mahalanobis distance. The aim is to find a way to balance these two distances in order to match the donors that have small values on both metrics.



For each of the combinations considered, performance will be evaluated in terms of prediction accuracy. R version 4.1.1 (2021-08-10)⁴ will be used to perform the analyses. The preferred journal for publication is Statistics in Medicine. In addition to the simulation study, the blended distance measures will be applied to empirical data from the SMOCC study.² These consist of the anonymised individual growth data of 1,933 children aged 0–15 months, as well as covariates that influence growth. Approval by the FETC has been obtained.

References (including additional references)

1. Van Buuren S. Curve matching: A data-driven technique to improve individual prediction of childhood growth. *Annals of Nutrition and Metabolism*. 2014;65(2-3):227-233. doi:10.1159/000365398
2. Herngreen WP, Van Buuren S, Van Wieringen JC, Reerink JD, Verloove-Vanhorick SP, Ruys JH. Growth in length and weight from birth to 2 years of a representative sample of Netherlands children (born in 1988–89) related to socioeconomic status and other background characteristics. *Annals of human biology*. 1994;21(5):449-463. doi:10.1080/03014469400003472
3. Buuren S van. Broken stick model for irregular longitudinal data. *Journal of Statistical Software*. 2020;Submitted for publication:1-47.
4. Team RC. R: A Language and Environment for Statistical Computing. Published online 2021. <https://www.R-project.org/>
5. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*. 2019;38(11):2074-2102. doi:10.1002/sim.8086
6. Ashworth A, Shrimpton R, Jamil K. Growth monitoring and promotion: Review of evidence of impact. *Maternal & child nutrition*. 2008;4:86-117. doi:10.1111/j.1740-8709.2007.00125.x.
7. Berkey CS, Kent RL. Longitudinal principal components and non-linear regression models of early childhood growth. *Annals of human biology*. 1983;10(6):523-536. doi:10.1080/03014468300006751
8. Van Dommelen P, Van Buuren S. Methods to obtain referral criteria in growth monitoring. *Statistical methods in medical research*. 2014;23(4):369-389. doi:10.1177/0962280212473301
9. Onis M de, Onyango AW. WHO child growth standards. *The Lancet*. 2008;371(9608):204. doi:10.1016/S0140-6736(08)60131-2
10. De Onis M, Onyango A, Borghi E, Siyam A, Blössner M, Lutter C. Worldwide implementation of the WHO child growth standards. *Public health nutrition*. 2012;15(9):1603-1610. doi:10.1017/S136898001200105X
11. De Onis M, Wijnhoven TM, Onyango AW. Worldwide practices in child growth monitoring. *The Journal of pediatrics*. 2004;144(4):461-465. doi:10.1016/j.jpeds.2003.12.034
12. Wilde JA de, Dommelen P van, Buuren S van, Middelkoop BJ. Height of South Asian children in the Netherlands aged 0–20 years: Secular trends and comparisons with current Asian Indian, Dutch and WHO references. *Annals of human biology*. 2015;42(1):38-44. doi:10.3109/03014460.2014.926988
13. Hu Y, He X, Tao J, Shi N. Modeling and prediction of children's growth data via functional principal component analysis. *Science in China Series A: Mathematics*. 2009;52(6):1342-1350. doi:10.1007/s11425-009-0088-5
14. Hauspie RC, Cameron N, Molinari L. *Methods in Human Growth Research*. Vol 39. Cambridge University Press; 2004.
15. Ng K, Sun J, Hu J, Wang F. Personalized predictive modeling and risk factor identification using patient similarity. *AMIA Summits on Translational Science Proceedings*. 2015;2015:132.
16. Schönbeck Y, Van Dommelen P, HiraSing RA, Van Buuren S. Trend in height of Turkish and Moroccan children living in the Netherlands. *PLoS One*. 2015;10(5):e0124686. doi:10.1371/journal.pone.0124686
17. Straatmann VS, Pearce A, Hope S, et al. How well can poor child health and development be predicted by data collected in early childhood? *J Epidemiol Community Health*. 2018;72(12):1132-1140.
18. Zimmerman DL, Núñez-Antón V, Gregoire TG, et al. Parametric modelling of growth curve data: An overview. *Test*. 2001;10(1):1-73. doi:10.1007/BF02595823
19. Efron B, Hastie T. *Computer Age Statistical Inference*. Vol 5. Cambridge University Press; 2016.
20. Siddique J, Belin TR. Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine*. 2008;27(1):83-102. doi:10.1002/sim.3001