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:

Date: 15 October 2021

1. Introduction

1.1. Growth prediction

Growth curve modeling is used to predict the future development of a child. It can provide answers to questions parents, physicians, and insurance companies may have concerning the future development of the child, the certainty of the future growth of the child, normal and healthy development, the certainty of prognosis, and effectiveness of treatment [1]. Curve matching [1] 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 children (usually 5) that are most similar to the target child. These are the so-called "people like me". In order to do this, we first need to match the children to the target child. The key question is: How do we obtain good matches? The current approach uses predictive mean matching (PMM). PMM consists of the following steps. First, a linear regression model is derived from the donor database. Then, this model is used to predict the values for all donors in the database and for the target child at a certain point in the future, for example at 14 months. Finally, the 5 donors which have the closest predicted value at 14 months are selected as the best matches. Their growth curves are then plotted and point estimates can be calculated by averaging the measurements.

1.2 Shortcomings of PMM

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 PMM and investigate alternative approaches. First, users of curve matching may solely be interested in how the future growth of a child would evolve, and therefore may find it difficult to select a future time point to base the matches on. In that case, we would rather be interested in the predictability over a period involving different ages. PMM is unlikely to be best on such a criterion. Second, the "optimal distance" that PMM creates may make the matches look unconvincing. Since different profiles may lead to the same predicted value, the values on separate predictors of some matches may be quite far from those of the target individual. Therefore, some users of curve matching question whether matches based on only future similarity are actually good matches, and whether historic similarity should be taken into account as well.

1.3 Objective

For the aforementioned reasons, it may prove useful to investigate alternative methods. In addition, PMM could be combined with these alternative methods by creating a "blended distance" measure. However, it is expected that this may come at the cost of predictive power. Therefore, the objective of this study is to investigate what the properties of a blended distance measure would be.

2. Strategy

2.1 Method

This study consists of simulation research and will report this according to the guidelines proposed by Morris, White, and Crowther [2]. Therefore, the planning of the study will follow the ADEMP structure, namely: Aims, Data-generating mechanisms, Estimands, Methods, and Performance measures. The methods that are considered will make use of the Mahalonobis distance, unsupervised KNN matching, and the Frechet distance. Then, the use of a blended distance measure will be investigated. For this blended distance, it will be evaluated what the influence is of varying the blending factor, the strength of the relationship between y and the first principal components of X, pre-selection on similarity, and the boosting of units. Performance

will be evaluated in terms of bias and coverage. R version 4.0.2 (2020-06-22) [3] will be used to perform the analyses. The candidate journal is Statistics in Medicine.

Note: What will the following be?

Data-generating mechanisms

Estimands

Methods

Performance measures

2.2 Data

In addition to the simulation study, empirical data from the SMOCC donor database will be used. This database contains individual growth data of 1,933 children aged 0–15 months, as well as covariates that influence growth. The study has been approved by the FETC.

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.

2. 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

3. Team R. RStudio: Integrated Development Environment for R. Published online 2020. http://www.rstudio.com/

4. Ashworth A, Shrimpton R, Jamil K. Growth monitoring and promotion: Review of evidence of impact. *Maternal & child nutrition*. 2008;4:86-117.

5. 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.

6. Van Dommelen P, Van Buuren S. Methods to obtain referral criteria in growth monitoring. *Statistical methods in medical research*. 2014;23(4):369-389.

7. Onis M de, Onyango AW. WHO child growth standards. *The Lancet*. 2008;371(9608):204.

8. 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.

9. De Onis M, Wijnhoven TM, Onyango AW. Worldwide practices in child growth monitoring. *The Journal of pediatrics*. 2004;144(4):461-465.

10. 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.

11. 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.

12. Meigen C, Hermanussen M. Automatic analysis of longitudinal growth data on the website willi-will-wachsen. De. *Homo*. 2003;54(2):157-161.

13. Hauspie RC, Cameron N, Molinari L. *Methods in Human Growth Research*. Vol 39. Cambridge University Press; 2004.

14. 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.

15. 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.

16. 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.

17. Zimmerman DL, Núñez-Antón V, Gregoire TG, et al. Parametric modelling of growth curve data: An overview. *Test.* 2001;10(1):1-73.

18. Efron B, Hastie T. *Computer Age Statistical Inference.* Vol 5. Cambridge University Press; 2016.