Research Master’s programme: Methodology and Statistics for the Behavioural, Biomedical and Social Sciences
Utrecht University

Research Report Anaïs Fopma (6199356)
Title: A blended distance to define "people-like-me"
Date: 19-12-2021

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

Preferred journal of publication: Statistics in Medicine
Word count:

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH REPORT

A blended distance to define "people-like-me"

Anaïs M. Fopma*

[1]Methodology & Statistics, Utrecht University, the Netherlands

Correspondence
*A.M. Fopma, Padualaan 14, 3584 CH Utrecht,the Netherlands. Email: a.m.fopma@uu.nl

Present Address
Padualaan14,3584CH Utrecht, the Netherlands

Summary

To be added later.

KEYWORDS:
curve matching, predictive mean matching, distance measures, metrics

1 | INTRODUCTION

The first three years of childhood form a crucial stage in determining children's subsequent development and health outcomes.[1] For this reason, growth monitoring is considered to be an integral part of paediatrics. It can aid in the identification of problems in development such as growth stunting, and ensure timely treatment or intervention to improve the child's health.[2] However, growth monitoring solely provides insights in the past and current developmental stages of the child. Growth curve modelling, on the other hand, can be used to predict future development. It can therefore provide more specific answers to questions health professionals, parents, and insurance companies may have, such as: 'Given what I know of the child, how will it develop in the future?' and 'Does this child get the most effective treatment among all options?'[3]

1.1 | Curve matching

An approach currently used for growth curve modelling is that of curve matching. Curve matching[3] is a Nearest Neighbour technique for individual prediction that constructs a prediction by aggregating the histories of "people like me". Its aim is to predict the growth of a target child by using the data of other children that are most similar to the target child. The donors that are selected as most similar to the target child are the so-called "people like me."

In order to do select the donors, similarity needs to be defined to match the donors to the target child. Therefore, the key question is: How are good matches obtained? The current approach uses predictive mean matching (PMM). PMM makes use of an existing donor database, containing the growth data of children that are older than the target child, and of which information at a later age is available. The first step is to fit a linear regression model on the donor database. Then, this model is used to predict the values for all donors 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 – usually five - 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. The growth patterns of the matched children suggest how the current child might evolve in the future.

An advantage of this technique is the high prediction accuracy.[3] Children of different ethnicities, for example, will differ in their growth.[4,5] Besides this, the matching approach offers more accurate information than using all data, which would predict the mean of the donor population irrespective of the person's growth. Moreover, the applications of curve modelling can be extended to settings other than the prediction of child health, such as patient recovery after an operation, prediction of longevity, and decision-making when multiple treatments are available.[3]

1.2 | Alternative approach

Even though PMM has proven to be promising in growth curve matching,[3] there are two reasons to move beyond the predictive distance used in PMM and investigate alternative metrics. Firstly, PMM requires users of curve matching to select a particular future time point to base the matches on (e.g. 14 months of age). In some cases, it may be difficult to choose this time point, especially when the 'future' is more vaguely defined as a time interval.[3] Secondly, the predictive distance may make the matches look unconvincing. The trajectories of the selected donors may all be close to the prediction for the target child at 14 months, but this does not imply that the histories are identical. After all, different profiles may lead to the same predicted value. Consequently, the curves of some of the matches may be quite far from the curve of the target child. Some users of curve matching feel that such discrepancies are undesirable, as these matches do not appear to be "people-like-me".

For these reasons, the practical implementation and use of curve matching can be improved by combining the predictive distance with alternative metrics that take into account historical similarity, thus creating a "blended distance" measure. Alternative metrics that quantify the difference between curves are, for example, the Mahalanobis distance and the Fréchet distance. It is expected that including these alternatives in a blended metric will come at the cost of predictability, while it will gain in proximity between curves. Therefore, the objective of this study is to investigate where to strike a balance between the predictive distance and alternative metrics.

2 | METHODS

This study consist of both simulation research and an application to empirical data. The data, metrics, and performance measures are discussed below.

2.1 | Broken stick model

The data of interest in this study are growth data of children and covariates that can be used to predict growth. These types of data consist of the height measurements of children at different observation times. It is important to note that the actual time points of data collection can sometimes differ substantially from the scheduled times. This may be due to a doctor's visit being planned during a holiday, the subject not showing up at the appointment, or the measurement device being out of order at the time of the scheduled observation.[6] As a consequence, the observation times will vary across subjects, and are said to be irregular. Irregular observation times present significant challenges for quantitative analysis. For example, it becomes more complex to predict the future from past data. Usually, a linear mixed model with time-varying outcomes is applied. An alternative is offered by the broken stick model[6], which converts irregularly observed data into a set of repeated measures. As a result, each child's growth trajectory can be approximated by a series of connected straight lines. The breakpoints between these lines are set to be the pre-specified, scheduled observation times. The advantage is that repeated measures data offers a lot more simplicity than the use of linear mixed models. Therefore, the *brokenstick* package[7] will be used for estimating the growth models in this study.

2.2 | Simulated data

For the simulation part of the study, the data-generating mechanisms will vary in their variance covariance structures. This means that we will sample different simulated data sets from populations consisting of data with different correlations. The correlations are taken to be 0.3, 0.5, and 0.7, respectively.

2.3 | Empirical data

After the simulation study is conducted, the different metrics to be evaluated will be applied to empirical data from the *Sociaal Medisch Onderzoek Consultatiebureau Kinderen* (SMOCK) study.[8] The SMOCK donor database contains the anonymised growth data of 1,933 children aged 0-15 months. In addition, the database contains covariates that influence growth, such as the sex, gestational age, birth weight, and height of the father and mother.

2.4 Metrics

In total, five metrics will be applied to the data: the predictive distance and four blended distance measures. The blended distance measures will consist of a combination of the predictive distance and the Mahalanobis distance, Fréchet distance, Hamming distance, and a locally supervised metric learning (LSML) measure, respectively. To create each blended distance, the following steps will be taken:

1. The predictive distance and the alternative distance are calculated for each donor in the database.
2. The two metrics are standardised.
3. The two metrics are plotted against each other to check if there is a linear relationship between them. If this is not the case, it will be investigated which transformation is necessary.

After the metrics have been calculated, the k most similar donors will be selected. For the predictive distance, the donors that will be selected are simply the ones with the smallest predictive distance. The *mice* package will be used to calculate this with PMM. For the blended metrics, the aim is to select the donors with a low value for the two distance measures that make up a blended metric. There are different possible approaches that can be taken to select these:

1. Create a linear combination with the two metrics, where both are assigned a weight: w1 * predictive distance + w2 * alternative distance. We then sort the donors based on this linear combination and select k donors with the lowest values.
2. First, calculate the squared deviation between each donor and the donor with the lowest value on the predictive distance. This is SD1. Then, calculate the squared deviation between each donor and the donor with the lowest value on the alternative distance. This is SD2. The donors with the smallest sum of SD1 and SD2 are selected as the best donors.
3. Use classification techniques to create a donor set, such as K-Nearest Neighbours. We select K donors that have the lowest values on both metrics, and are thus closest to the origin.
4. Use a non-linear decision boundary to select the donors that lie closest to the origin.

As an example, the first combination of the predictive distance and the Mahalanobis distance is illustrated in Figure 1. Here, the data of 200 children from the SMOCK study are used. The first subject is taken as the target, the 199 other subjects as the donors. For all donors, the Mahalanobis distance for the measurements during the first six months of growth is calculated. In addition, the predictive distance between each donor and the target is calculated. In the figure, the Mahalanobis distance and predictive distance are plotted against each other. The red donors are the five matches with the smallest predictive distance, where especially subject 10006 has a large Mahalanobis distance. The triangular donors are the five matches with the lowest Mahalanobis distance, where especially subject 11018 has a large predictive distance. A blended distance would balance the distance measures, such that the donors with a low value for both distance measures are chosen. These are circled in green.

Each of the possible blended metrics are discussed in the following sections.

2.4.1 Blended metric 1: combination with Mahalanobis distance

The Mahalanobis distance is defined as the distance between two N dimensional points scaled by the statistical variation in each component of the point. For example, if $\vec{x}$ and $\vec{y}$ are two points from the same distribution which has covariance matrix $\mathbf{C}$, then the Mahalanobis distance is given by

$$((\vec{x} - \vec{y})' \mathbf{C}^{-1} (\vec{x} - \vec{y}))^{\frac{1}{2}}$$

2.4.2 Blended metric 2: combination with Fréchet distance

The Fréchet distance is a measure of similarity between two curves. It is commonly described by the following analogy:

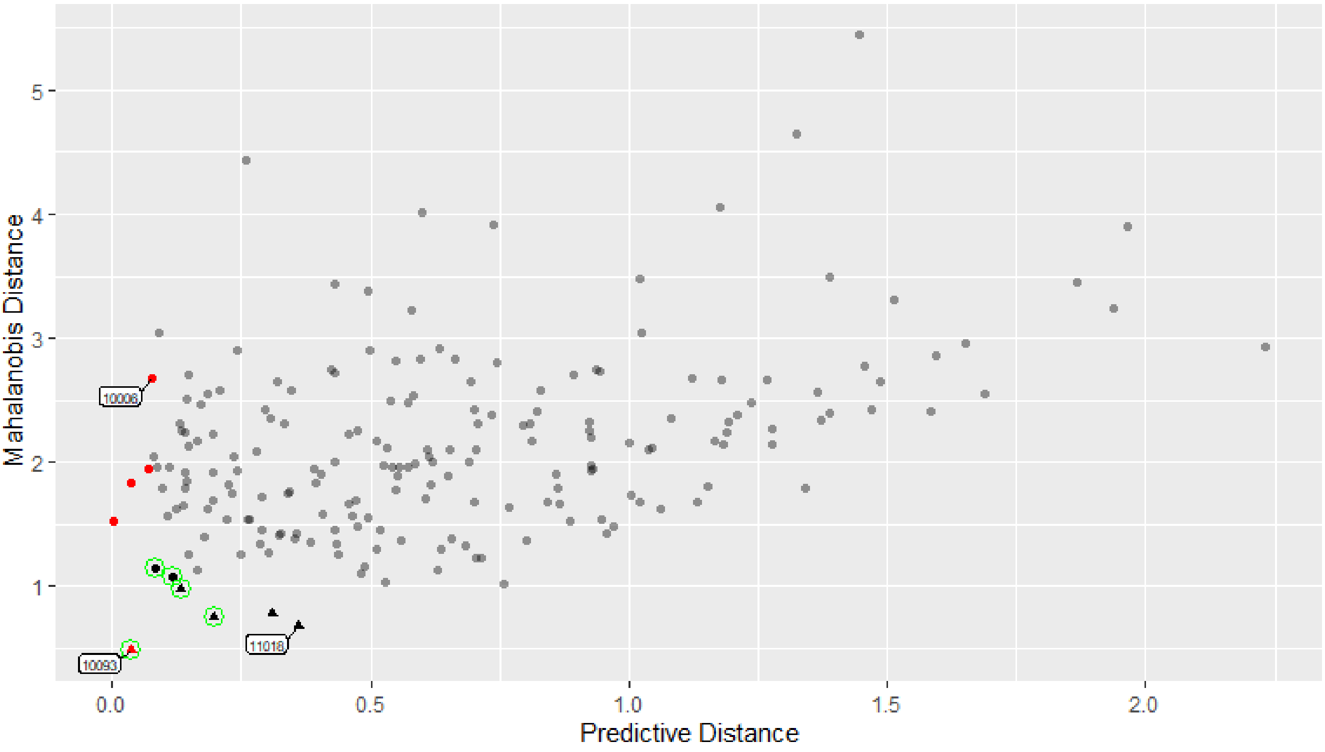

FIGURE 1 Mahalonobis distance plotted against predictive distance for each of the 199 donor children.

> A man is walking a dog on a leash: the man can move on one curve, the dog on the other; both may vary their speed, but backtracking is not allowed. What is the length of the shortest leash that is sufficient for traversing both curves?[9]

The Fréchet distance takes into account the location and ordering of the points along the curves.[9] The definition is the following.[9,10] We define a curve as a continuous mapping $f : [a, b] \rightarrow V$, where $a, b \in< \mathfrak{R}$ and $a \leq b$ and (V, d) is a metric space. Given two curves $f : [a, b] \rightarrow V$ and $g : [a', b'] \rightarrow V$, their Fréchet distance is defined as

$$\delta_F(f, g) = \inf_{\alpha,\beta} \max_{t\in[0,1]} d(f(\alpha(t)), g(\beta(t)))$$

where α (resp. β) is an arbitrary continuous nondecreasing function from [0, 1] onto $[\alpha, \beta]$ (resp. $[a', b']$).

2.4.3 | Blended metric 3: combination with Hamming distance

The Hamming distance between two equal-length strings of symbols is the number of positions at which the corresponding symbols are different.[11]

2.4.4 | Blended metric 4: combination with LSML measure

Locally Supervised Metric Learning (LSML) is a trainable similarity measure that is used to identify a cohort of patients from a training set that are most clinically similar to a test patient.[12] The distance metric is defined as

$$D_{LSML}(x_i, x_j) = \sqrt{(x_i - x_j)^T W W^T (x_i - x_j)}$$

where x_i and x_j are the patient feature vectors for patients i and j respectively and W is a transformation matrix that is estimated from the training data.

TABLE 1 Example display of results

Correlation	Metric	Predictability	Proximity
0.3	Predictive distance		
	Blended metric 1		
	Blended metric 2		
	Blended metric 3		
	Blended metric 4		
0.5	Predictive distance		
	Blended metric 1		
	Blended metric 2		
	Blended metric 3		
	Blended metric 4		
0.7	Predictive distance		
	Blended metric 1		
	Blended metric 2		
	Blended metric 3		
	Blended metric 4		

2.5 | Performance measures

As the objective is to match a number of donors and estimate with those donors the growth of the target child, it is necessary to evaluate how well the donors actually match the target. In order to do this, performance will be measured in two ways: in terms of proximity and predictability. Proximity is defined as the closeness of the curves of the selected donors to the curve of the target. Predictability is defined as the extent to which the selected donors predict the growth of the target, i.e. the explained variance.

We consider a full-factorial simulation study, where all possible combinations of metrics and data-generating mechanisms are evaluated. R version 4.1.1 (2021-08-10)[13] will be used to simulate the data and perform the analyses.

3 | RESULTS

In Table 1, a simple example is provided of how the results of the simulation study could be presented.

References

1. 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. Publisher: BMJ Publishing Group Ltd.

2. Cordeiro JR, Postolache O, Ferreira JC. Child's target height prediction evolution. *Applied Sciences* 2019; 9(24): 5447. Publisher: Multidisciplinary Digital Publishing Institute.

3. 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. Publisher: Karger Publishersdoi: 10.1159/000365398

4. Wilde dJA, Dommelen vP, Buuren vS, 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. Publisher: Taylor & Francisdoi: 10.3109/03014460.2014.926988

5. 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. Publisher: Public Library of Science San Francisco, CA USAdoi: 10.1371/journal.pone.0124686

6. Buuren vS. Broken stick model for irregular longitudinal data. *Journal of Statistical Software* 2020; Submitted for publication: 1–47.

7. Buuren vS. Broken Stick Model for Irregular Longitudinal Data. .

8. 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. Publisher: Taylor & Francisdoi: 10.1080/03014469400003472

9. Eiter T, Mannila H. Computing discrete Fréchet distance. tech. rep., Citeseer; 1994.

10. Alt H, Godau M. Measuring the resemblance of polygonal curves. In: ; 1992: 102–109.

11. Waggener B, Waggener WN, Waggener WM. *Pulse code modulation techniques*. Springer Science & Business Media . 1995.

12. 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. Publisher: American Medical Informatics Association.

13. Team RC. R: A Language and Environment for Statistical Computing. 2021.

14. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine* 2019; 38(11): 2074–2102. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8086doi: 10.1002/sim.8086

15. Ashworth A, Shrimpton R, Jamil K. Growth monitoring and promotion: review of evidence of impact. *Maternal & child nutrition* 2008; 4: 86–117. Publisher: Wiley Online Librarydoi: 10.1111/j.1740-8709.2007.00125.x.

16. 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. Publisher: Taylor & Francisdoi: 10.1080/03014468300006751

17. Van Dommelen P, Van Buuren S. Methods to obtain referral criteria in growth monitoring. *Statistical methods in medical research* 2014; 23(4): 369–389. Publisher: Sage Publications Sage UK: London, Englanddoi: 10.1177/0962280212473301

18. Onis dM, Onyango AW. WHO child growth standards. *The Lancet* 2008; 371(9608): 204. Publisher: Elsevierdoi: 10.1016/S0140-6736(08)60131-2

19. 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. Publisher: Cambridge University Pressdoi: 10.1017/S136898001200105X

20. De Onis M, Wijnhoven TM, Onyango AW. Worldwide practices in child growth monitoring. *The Journal of pediatrics* 2004; 144(4): 461–465. Publisher: Elsevierdoi: 10.1016/j.jpeds.2003.12.034

21. 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. Publisher: Springerdoi: 10.1007/s11425-009-0088-5

22. Hauspie RC, Cameron N, Molinari L. *Methods in human growth research*. 39. Cambridge University Press . 2004.

23. Zimmerman DL, Núñez-Antón V, Gregoire TG, et al. Parametric modelling of growth curve data: An overview. *Test* 2001; 10(1): 1–73. Publisher: Springerdoi: 10.1007/BF02595823

24. Efron B, Hastie T. *Computer age statistical inference*. 5. Cambridge University Press . 2016.

25. Siddique J, Belin TR. Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine* 2008; 27(1): 83–102. Publisher: John Wiley & Sons, Ltddoi: 10.1002/sim.3001

26. Buuren vS, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 2011; 45(3): 1–67.

27. Alt H, Godau M. Computing the Fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications* 1995; 5(01n02): 75–91. Publisher: World Scientific.

How to cite this article: Fopma A.M (2022), A blended distance to define "people-like-me", *Statistics in Medicine*, *x;x:x–x*.