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: 2192

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

Curve matching is a powerful tool to predict the development of a child (target) with the data of other children (donors). The current technique relies on predictive mean matching, which matches donors that are most similar to the target based on the predictive distance. Even though this approach leads to high prediction accuracy, there are two disadvantages. Firstly, it requires users of curve matching to select a particular future time point to base the matches on. In some cases, it may be difficult to choose this time point. Secondly, the predictive distance may make matches look unconvincing, as the profiles of the matched donors can substantially differ from the profile of the target, even if they are close on the predicted time point. To counterbalance these disadvantages, similarity between the curves of the donors and the target can be taken into account when selecting donors. The objective of the current study is to do so by combining the predictive distance measure with alternative metrics that focus on similarity between the profiles of the subjects, thus creating a 'blended distance' measure. This blended metric is evaluated in terms of proximity and predictability.

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.

In order to select these 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 thus suggest how the target child might evolve in the future.

An advantage of this technique is its high prediction accuracy.[3] Moreover, the applications of curve matching can be extended to settings other than the prediction of child development, 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.[4] 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. Examples of these alternative metrics that quantify the difference between curves are 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, such that donors are selected that are both similar to the target in profile, and predict its future development accurately.

2 | METHODS

This study consist of both simulation research and an application to empirical data. The data, metrics to be evaluated, 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 will 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.[4] 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, as it becomes more complex to predict the future from past data. Usually, a linear mixed model with time-varying outcomes is applied. However, an alternative is offered by the broken stick model,[4] 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 offer a lot more simplicity than the use of linear mixed models. Therefore, the *brokenstick* package[5] will be used for estimating the growth models in this study.

2.2 | Simulated data

For the simulation part of the study, all data are generated from a three populations consisting of data with varying correlations. The correlations are taken to be 0.3, 0.5, and 0.7, respectively. For each population, we define 5 continuous predictor variables, corresponding to the standardised height measurements (Z-scores) at birth, the age of 1 month, 2 months, 3 months, and 6 month, respectively. We define 1 continuous outcome, corresponding to the predicted standardised height measurement at the age of 14 months. In each simulation iteration, we draw two samples from the population: a target set (n = 1) and a donor set (n = 1000). The data generating mechanism of the predictor space is a multivariate normal distribution, $\mathbf{X} = \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where mean vector $\boldsymbol{\mu} = [0, 0, ..., 0]$. The covariance matrix Σ for the first population with a correlation of 0.3 is given by:

$$\Sigma = \begin{bmatrix} 1.03 & 0.30 & 0.36 & 0.33 & 0.28 \\ 0.30 & 0.94 & 0.26 & 0.34 & 0.25 \\ 0.36 & 0.26 & 1.02 & 0.34 & 0.26 \\ 0.33 & 0.34 & 0.34 & 1.12 & 0.27 \\ 0.28 & 0.25 & 0.26 & 0.27 & 0.94 \end{bmatrix}$$

The covariance matrix Σ for the first population with a correlation of 0.5 is given by:

$$\Sigma = \begin{bmatrix} 1.07 & 0.58 & 0.53 & 0.51 & 0.53 \\ 0.58 & 1.05 & 0.53 & 0.46 & 0.49 \\ 0.53 & 0.52 & 0.99 & 0.49 & 0.52 \\ 0.51 & 0.46 & 0.49 & 0.97 & 0.51 \\ 0.53 & 0.49 & 0.52 & 0.51 & 1.03 \end{bmatrix}$$

The covariance matrix Σ for the first population with a correlation of 0.7 is given by:

$$\Sigma = \begin{bmatrix} 0.97 & 0.65 & 0.67 & 0.68 & 0.68 \\ 0.65 & 0.98 & 0.66 & 0.70 & 0.66 \\ 0.67 & 0.66 & 0.97 & 0.71 & 0.68 \\ 0.68 & 0.70 & 0.71 & 1.05 & 0.71 \\ 0.68 & 0.66 & 0.68 & 0.71 & 0.97 \end{bmatrix}$$

I have taken the example Hanne gave me, and just filled in the information as it would be for this case. Does the description make sense like this?

Like Mingyang proposed, we can also vary the data in dimensionality. I will keep this in mind, but leave it out of the research report for now.

The spacing between the matrices is a bit ugly, but vspace only seems to work after a line of text. Any tips?

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.[6] 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, four 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, and a locally supervised metric learning (LSML) measure, respectively. I agree with Mingyang to first focus on the combination with the Mahalanobis distance, but I will still include the other combinations in the report for now To create each blended distance, the following steps will be taken:

1. The predictive distance and the alternative distance are calculated for each donor.

2. The two metrics are standardised.If this is done, is weighting still necessary? Or would weighting be included for the purpose of creating a blending factor that can be manipulated by the user, like Stef said in the project description: Potentially, the end user can manipulate the blending factor in order to place differential emphasis on "historic similarity" versus "future similarity". If we do look into weighting, should I include it in the report (e.g. say that we will look into a blending factor of 0.25, 0.5, 0.75, or leave it out for now?

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*[7] 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 the blended metric. In order to do so, the blended metric D_B is calculated for each donor as the absolute sum of the predictive distance D_1 and the alternative distance D_2 Is it necessary to say absolute, or is this redundant as the values will always be positive?:

$$D_B = |D_1 + D_2|.$$

Or, if we include weighting: In order to do so, the blended metric D_B is calculated for each donor as a linear combination of the predictive distance D_1 and the alternative distance D_2:

$$D_B = w_1 D_1 + w_2 D_2,$$

where w_1 is the weight assigned to the predictive distance and w_2 is the weight assigned to the alternative distance.

Or, as Stef proposed: In order to do so, the rank is calculated for the predictive distance D_1 and the alternative distance D_2, where ties are randomly broken. The blended distance D_B is then given by:

$$D_B = p * rank_1 + (1 - p) * rank_2,$$

where $0 \leq p \leq 1$.

The k donors with the lowest values on D_B are selected as the best matches.

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:

> 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?[8]

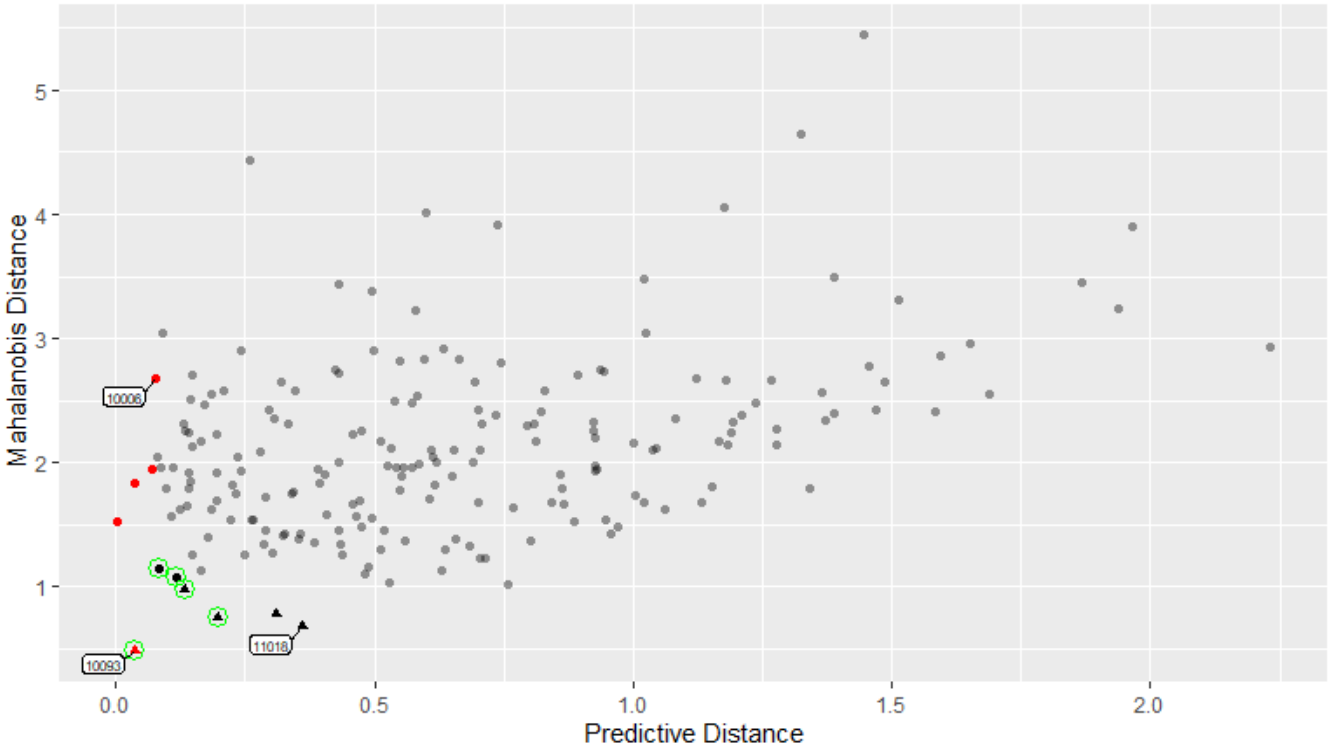

FIGURE 1 Mahalanobis distance plotted against predictive distance for each of the 199 donors.

The Fréchet distance takes into account the location and ordering of the points along the curves.[8] The definition is the following.[8,9] We define a curve as a continuous mapping $f : [a, b] \to V$, where $a, b \in \mathfrak{R}$ and $a \leq b$ and (V, d) is a metric space. Given two curves $f : [a, b] \to V$ and $g : [a', b'] \to V$, 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 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.[10] The distance metric is defined as

$$D_{LSML}(x_i, x_j) = \sqrt{(x_i - x_j)'WW'(x_i - x_j)},$$

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.

2.5 | Performance measures

As the objective of this study is to investigate where to strike a balance between the predictive distance and alternative metrics, it is necessary to evaluate how well the selected donors 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. It will be measured by the the sum of squared differences of the broken stick estimates, given by:

$$SS = \sum_{i,j=1}^{n} (y_{i,j} - x_i)^2,$$

TABLE 1 Example display of synthetic estimates of performance for the measures of interest.

Correlation	Metric	R^2	SS
0.3	Predictive distance	0.94	92.18
	Blended metric 1	0.83	77.88
	Blended metric 2	0.77	79.04
	Blended metric 3	0.65	86.66
0.5	Predictive distance	0.96	93.43
	Blended metric 1	0.76	56.42
	Blended metric 2	0.88	66.78
	Blended metric 3	0.55	54.49
0.7	Predictive distance	0.88	97.42
	Blended metric 1	0.73	47.99
	Blended metric 2	0.74	52.90
	Blended metric 3	0.78	51.09

where x_i is the i^{th} broken stick estimate of the target and y_i is the i^{th} broken stick estimate of donor j. I'm not so sure about this notation, is it correct? This performance measure will give an indication of the similarity between the profiles of the donors and the target. Predictability is defined as the extent to which the selected donors predict the growth of the target. It will be measured by the explained variance, So is this the proportion of variance in the target explained by the k selected donors? given by:

$$R^2 = 1 - \frac{SSR}{SST}.$$

For the explained variance:

- Do we take the SSR as the difference between the curve of the target and the average of the curves of the k donors? So if we compare it to regression, in this case the average of the curves of the k donors would be used in place of what is normally the regression line?
- And do we take the SST as the difference between the curve of the target and the mean of the y values of the target?

This performance measure will give an indication of how well the selected donors predict the future development of the target.

I had this question before: Would it also be useful to include as a performance measure the bias, that is, the bias between the value that is predicted by the k donors at e.g. 14 months and the actual value for the target at 14 months? If I understand correctly, this is what you mentioned in the meeting. Should R squared then be replaced by the RMSD? Something like this? Predictability is defined as the extent to which the selected donors predict the growth of the target. It will be measured by the root-mean-square deviation (RMSD), given by:

$$RMSD = \sqrt{\frac{\Sigma_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}},$$

where x_i is the actual value for the target at 14 months, $\hat{x}_i$ is the value at 14 months predicted by the k donors, and N is the number of simulations.

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)[11] 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. The results given are synthetic.

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

5. Van Buuren S. Broken Stick Model for Irregular Longitudinal Data. .

6. 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

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

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

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

10. 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.

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

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