Data-driven donation strategies: understanding and predicting blood donor deferral

Vinkenoog, M.

CHAPTER

5

Challenges and limitations in clustering blood donor hemoglobin trajectories

Published in: International workshop on advanced analysis and learning on temporal data, 72-84. doi:10.1007/978-3-030-39098-3_6

Authors: M Vinkenoog, MP Janssen, M van Leeuwen

Abstract

Background - In order to prevent iron deficiency, Sanquin measures a blood donor's hemoglobin level before each donation and only allows a donor to donate blood if it is above a certain threshold. In around 6.5% of blood bank visits by women, the donor's hemoglobin level is too low and the donor is deferred from donation. For visits by men, this occurs in 3.0% of cases. To reduce the deferral rate and keep donors healthy and motivated, we would like to identify donors that are at risk of having a low hemoglobin level. To this end we have historical hemoglobin trajectories at our disposal, i.e., time series consisting of hemoglobin measurements recorded for individual donors.

Methods - As a first step towards our long-term goal, in this paper we investigate the use of time series clustering. Unfortunately, existing methods have limitations that make them suboptimal for our data. In particular, hemoglobin trajectories are of unequal length and have measurements at irregular intervals. We therefore experiment with two different data representations. That is, we apply a direct clustering method using dynamic time warping, and a trend clustering method using model-based feature extraction. In both cases the clustering algorithm used is k-means.

Results - Both approaches result in distinct clusters that are well-balanced in size. The clusters obtained using direct clustering have a smaller mean within-cluster distance, but those obtained using the model-based features show more interesting trends. Neither approach results in ideal clusters though. We therefore conclude with an elaborate discussion on challenges and limitations that we hope to address in the near future.

Introduction

Sanquin is the national blood bank in the Netherlands. Every year, about 300 000 donors visit the blood bank, resulting in over 420 000 donations a year. Women are allowed to donate up to three times a year, men up to five times. There are many policies in place to ensure that the blood products that are collected are safe for the patients they will be given to. Moreover, Sanquin has the responsibility to prevent volunteer blood donors from developing health problems related to blood donation. One big risk of regular blood donation is anemia due to low iron stores or iron deficiency. A whole blood donation takes about 500 mL of blood from the donor, which contains 210 to 240 mg iron bound to hemoglobin. The total concentration of iron in the human body is approximately 38 mg/kg body weight for women and 50 mg/kg body weight for men, so a single blood donation constitutes a substantial loss of iron. [16, 17]

To prevent donors from becoming iron deficient, their hemoglobin levels are checked before each blood donation. Based on the hemoglobin measurement it is decided whether they may donate at that time: the lower limit for donation is 7.8 mmol/L for women, and 8.4 mmol/L for men. When a donor is below the threshold, they are sent home and can return for donation a few weeks later. This type of deferral occurs quite frequently: about 6.5% of female and 3.0% of male donors have too low hemoglobin levels when they visit the blood bank.

The large number of deferrals is problematic, both for donors and the blood bank: being deferred from donation is demotivating for the donor, who may decide not to return in the future, and not efficient for the blood bank, leading to a higher cost per blood product.

Because of this, Sanquin and other blood banks internationally spend considerable resources on investigating ways to reduce the deferral rate while keeping donors healthy. One asset that can be exploited for this are the hemoglobin measurements that blood banks have recorded in the past. In this paper we report on a preliminary study investigating whether we can distinguish groups of donors having different trends in their hemoglobin trajectories; if this is the case, these trends could be used to devise more personalised invitation and deferral policies.

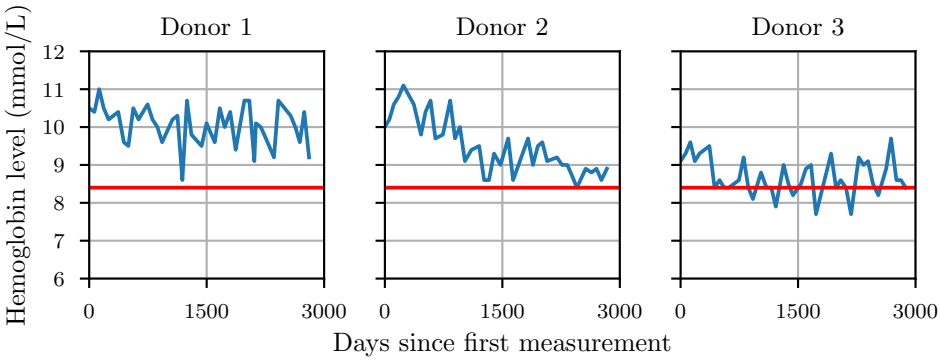

Figure 5.1: Hemoglobin trajectories of three male donors. From left to right: a high stable trajectory, a declining trajectory, and a low stable trajectory. The red line is the hemoglobin threshold for donation (8.4 mmol/L for male donors).

5

Approach and contributions

We have data available on all blood bank visits in the Netherlands since 2006. For every donor, we have only two relevant background variables: year of birth and sex. It has long been known that age and sex affect hemoglobin levels. Men's levels are higher than women's and decrease with age, while women's levels increase after menopause. [62]

Apart from these factors, a large part of the variation in hemoglobin levels can be attributed to diet and lifestyle: the iron richness of the donor's diet and their activity level play a substantial role here. However, we don't have large-scale data on this. The clusters of donors we hope to identify could be a proxy for these variables.

The more interesting part of the data are the hemoglobin measurements taken every time the donor visits the blood bank. Each measurement has a time stamp, and together the individual measurements of a single donor form a time series; we will refer to these time series as hemoglobin trajectories.

We aim to find groups of donors whose hemoglobin levels are similar throughout their donation history. More specifically, we would like to distinguish donors with a stable (high or low) hemoglobin level from donors with a declining level over time, as these require different donation policies. The three different trends that we expect to find in the data are illustrated in Figure 5.1.

Finding groups of similar data points in an unsupervised manner is a typical clustering task and since hemoglobin trajectories are time series, we naturally resort to

time series clustering. Time series clustering can be applied in many fields and has been studied for a long time, as a result of which a large number of clustering methods for time series exist. [63, 64, 65]

A big limitation, however, is that most existing algorithms require the time series to be sampled at fixed, equidistant time stamps. In our data, the sampling intervals are highly irregular on two levels. First, the intervals are not uniform across time series; an easy example is that women are allowed to donate three times a year, men five times. Second, the intervals are not uniform within the time series either: sometimes a donor returns for their next donation two months after the previous one, sometimes six months. Donors can also temporarily stop donating, and then return years later. A related limitation that is relevant to our data is that the time series have unequal lengths. Many donors in the data set have been regularly donating for over ten years, while others have just started.

Faced with these challenges, in this paper we will investigate whether we can transform our data for use with a standard clustering method without losing critical information. Specifically, we will employ two approaches:

1. Direct clustering using re-sampling combined with dynamic time warping [66] as distance measure;
2. Trend clustering using model-based feature extraction combined with the Euclidean distance.

As our main aim is to evaluate and compare the data representations, the choice of a clustering method is less important; we will use k-means because it is straightforward, effective, and well-known. [67]

The main contributions of our preliminary study are a proof-of-concept showing that clustering of hemoglobin trajectories of Dutch blood donors is feasible, and the identification of challenges and limitations of using time series clustering for hemoglobin trajectories. We consider these to be important first steps towards an effective clustering method for irregular time series in which the irregularities itself may contain useful information.

Data

Our data consists of all blood donations made at any of Sanquin's locations between January 2006 and June 2018, extracted from the blood bank's database system eProgesa. In total, there are 6 945 611 donations by 688 665 unique donors. Because we

are interested in donors' hemoglobin trajectories from their first donation onward, we selected for our analyses all donors that did not visit the blood bank before 2010. It is possible that there are donors in the data set that donated before 2006 and returned after a gap of at least four years, but we expect this number to be low, and their hemoglobin levels similar to actual new donors.

Many types of blood donation take place at Sanquin, the most common being plasma donation and whole blood donation. During plasma donation, red cells are returned to the donor and only the plasma is collected. As hemoglobin is contained in the red blood cells, this type of donation does not have a substantial effect on hemoglobin levels. Therefore, we only look at donors that donate whole blood, during which no blood components are returned to the donor.

We take into account donors that have donated whole blood at least eight times in our time window—once a year on average. There are 23 856 female and 20 299 male donors that fit these criteria. To decrease computation time, we randomly selected 5000 women and 5000 men for our experiments. Within this data set, the deferral rate due to low hemoglobin is 7.8% among female donors and 3.3% among male donors.

The two data sets contain 5000 individual univariate time series each, consisting of the hemoglobin measurements during the visits to the blood bank. Hemoglobin is measured in mmol/L. The median number of measurements per time series is 12 for women (interquartile range, IQR 10–14) and 14 for men (IQR 11–19).

The time intervals between measurements differ both within and between time series. The minimum required interval between two donations is 122 days for women and 56 days for men, but it can even be a few years. The median interval for women is 133 days (IQR 112–169) and for men 79 days (IQR 64–114). Aside from the hemoglobin measurements, the only variable used is the sex of the donor. Clustering methods will be applied separately to the female and male subsets.

Methods

We will experiment with two data representations and compare the results of the k-means clustering algorithm on both representations. The methods will be compared on cluster tightness using mean within-cluster distance, and visually on the informativeness of the cluster using the graphs of the cluster centroids.

The first method employs direct clustering using dynamic time warping based on the hemoglobin levels at each time point, the second method employs trend clustering using model-based feature extraction. Preprocessing is the same for both.

Preprocessing

When time series are of equal length and have the same measurement intervals, clustering is relatively straightforward. At each time point, we can calculate the difference between measurements in two time series, and group time series with smaller differences in the same cluster. However, from this perspective our data is rather messy: time series are all of differing lengths and have different measurement intervals, both within and between individuals. While there are more sophisticated ways to handle this (see the Discussion), none of the existing algorithms that we found are perfectly suited to our data. Therefore, for this first trial we decided to side-step the problem of unequal intervals by resampling the time series to regular intervals by linear interpolation.

We take each donor's first donation since 1 January 2010 as the starting point of their time series. All time stamps are relative to the first time stamp, recorded as days since first donation. Hemoglobin values are then resampled to weekly measurements using linear interpolation. This gives a maximum of 439 observations per donor, one for each week between 1 January 2010 and 1 June 2018. Donors that started donating later in the time window will have fewer measurements, and thus have a number of missing values at the tail of the time series. For the first 140 weeks, the number of donors with missing values is almost zero, but then the number of donors that still has measurements starts dropping at a steady rate. We chose to use hemoglobin measurements up to 286 weeks after the first donation, at which time half of our 5000 donors has no missing values, and the other half misses at most 50% of observations.

Direct clustering using dynamic time warping

For this method, the features that we will feed to the clustering algorithm are the resampled hemoglobin measurements as described in the previous section. As a distance measure, we use dynamic time warping (DTW) with the window parameter set to $w = 5$. [66] This algorithm is better-suited to our data than for instance the Euclidean distance, because it takes into account varying speeds and time shifts. Because the time series vary in length, we compare time series only up to the last data point in the shortest series.

The algorithm can be summarised as follows:

1. Calculate the Euclidean distance between the first point in the first series, and every point within the window of $w = 5$ in the second series;
2. Store the minimum distance calculated;
3. Repeat steps 1–2 for all points in the first series;
4. Add all the minimum distances to get the DTW distance.

Trend clustering using model-based feature extraction

The second method takes as input for the clustering algorithm not the (resampled) time series itself, but rather a set of features that should summarise the time series in such a way that similar time series will have similar feature values. We are interested in distinguishing three types of hemoglobin trajectories: high stable, low stable, and declining. We therefore choose to cluster the trajectories based on the intercept and slope of the linear trend.

The intercept and slope are calculated using linear least-squares regression on the resampled time series described in the previous section, to allow for a direct comparison between the two methods. Because the slope and intercept values are on different scales, we normalise them using a min-max scaler before clustering. The values are then all between 0 and 1, 0 being the minimum value among the time series and 1 the maximum.

Clustering algorithm

For the actual clustering, we use k-means clustering, a heuristic algorithm that is usually quite fast at finding a local optimum. [67] It requires the user to specify the number of desired clusters k. We chose this well-known algorithm for its wide applicability and straightforward implementation.

For the direct clustering, the input to the algorithm contains the resampled time series. Because of the differing lengths of the time series, we chose to initialise the clusters randomly from a uniform distribution, instead of choosing k time series as initial cluster centroids. The distance measure used is DTW.

For the trend clustering, the input consists of two features per trajectory: the intercept and the slope of the linear trend. As distance measure the Euclidean distance is used.

In general, k-means clustering returns the best results if the algorithm stops when the difference between the cluster centroids in two subsequent iterations is smaller than some ϵ. Because the program is computationally expensive due to the DTW calculations, we opted to let it run for at most five iterations for the first clustering method.

The algorithm is as follows:

1. Initialise k cluster centroids;
2. Assign each time series to the cluster to which it is most similar, based on the specified distance;
3. Recalculate the cluster centroids by taking the average value for each feature;
4. Repeat steps 2–3 for 5 iterations or until convergence.

Evaluation

We compare the clusters based on the two data representations in two ways: cluster tightness and cluster informativeness. The first is a numerical comparison, the second graphical. Cluster tightness is assessed by the mean within-cluster distance. For each cluster, we calculate the distance from the cluster centroid to the individual time series by taking the DTW distance between the two. The mean of these distances is the mean within-cluster distance. We also calculate the sum of the within-cluster distance for each value of k, which is the sum of the DTW distances between the individual time series and the cluster centroids, summed over all clusters. As the number of clusters increases, we expect the sum of the within-cluster distances to decrease.

Cluster informativeness is assessed visually by looking at the graphs of the cluster centroids. We hope to see centroids that are different in slope, and not just horizontal lines with different average hemoglobin values.

Results

We will first present the results from both methods separately, then compare the two on cluster tightness and informativeness.

Direct clustering

Figure 5.2 shows the centroids of the clusters after direct clustering with DTW. At $k = 2$ and $k = 3$, we see that the clusters are based mostly on the mean hemoglobin

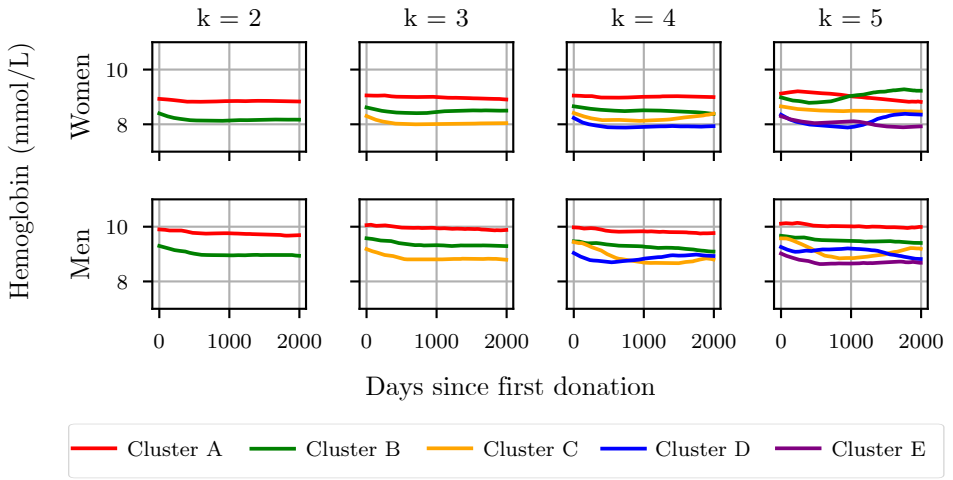

Figure 5.2: Cluster centroids after clustering resampled hemoglobin trajectories of 5000 female and 5000 male donors with the k-means clustering algorithm ($k \in [2, 3, 4, 5]$) and DTW distance as distance measure.

level in the donors, and cluster centroids are almost parallel. At higher numbers of clusters, we start to see some differences in trends as well, with centroids intersecting each other. At $k > 5$, we saw that centroids start overlapping for longer periods of time and are no longer distinct enough to be informative. These graphs are not included in the paper. In almost all centroids, there is a decrease in hemoglobin value at the beginning of the hemoglobin trajectory.

To assess the tightness of the clusters, Table 5.1 shows the mean within-cluster distances, with DTW used as distance measure. The total sum of the within-cluster distances decreases as the number of clusters increases, which is expected because the same distance measure was used to create the clusters. The names of the clusters correspond to those in Figure 5.2. Table 5.1 also shows the number of time series assigned to each cluster. We see that in size, the clusters are quite well-balanced: the smallest cluster has size 413 where size 1000 would be expected if all clusters were the same size (female donors, $k = 5$, cluster B).

Trend clustering

Figure 5.3 shows the cluster centroids after clustering on trend features. As after the direct clustering, the centroids are distinct from each other and do not intersect at $k = 2$ and $k = 3$. From $k = 4$ and up, cluster B shows an interesting new trend in

Sex	k	Cluster A $\bar{d}$ (N)	Cluster B $\bar{d}$ (N)	Cluster C $\bar{d}$ (N)	Cluster D $\bar{d}$ (N)	Cluster E $\bar{d}$ (N)	Sum $\bar{d}$
Women	2	7.1 (1613)	6.3 (3387)				32670
	3	5.7 (2128)	7.0 (986)	5.9 (1886)			30135
	4	5.4 (1197)	5.2 (1205)	5.9 (1671)	6.9 (927)		29049
	5	6.3 (475)	6.5 (413)	5.5 (1379)	5.9 (1766)	5.4 (967)	28840
Men	2	7.2 (2020)	6.3 (2980)				33260
	3	6.1 (1997)	5.6 (1851)	6.9 (1152)			30508
	4	7.1 (1600)	5.8 (1589)	5.5 (835)	5.1 (976)		30222
	5	4.9 (896)	5.9 (1424)	5.2 (906)	5.5 (871)	6.8 (903)	28567

Table 5.1: The mean distance from the centroid to the time series ($\bar{d}$) and the number of time series in each cluster (N) after direct clustering. Dynamic time warping is used as distance measure. The rightmost column shows the sum of the within-cluster distances.

male donors: the slope of the line is much steeper than those of the other clusters.

In Table 5.2, we see that the mean distance from the centroid to the individual time series is larger than in the clusters obtained using the first method. The sum of the within-cluster distances does not decrease as k increases, and for female donors it even increases substantially. This can happen because in this method, the clusters are decided based on the Euclidean distances between the trend features of the time series, rather than the DTW distance between time series as in the first method.

The number of time series per cluster is mostly well-balanced, although there are some cases of small clusters: at $k = 5$, in male donors, cluster A only contains 386 time series where 1000 would be expected if all clusters were of equal size.

Comparison

From the within-cluster distances, it is clear that the direct clustering method leads to tighter clusters. Figure 5.4 illustrates this well. It shows the result of both direct and trend feature clustering on male donors with $k = 4$ clusters. Each subplot shows the cluster centroid in red, and 100 randomly selected individual time series within the cluster in grey. Although after both direct and trend clustering the cluster centroid lies in the middle of the individual time series, the spread is much smaller in direct than in trend clustering.

In both methods, cluster centroids vary mostly in the average hemoglobin value over time, and not as much in trend, which is what we are mostly interested in. The exception is cluster B in the trend clustering method, which shows a relatively steep

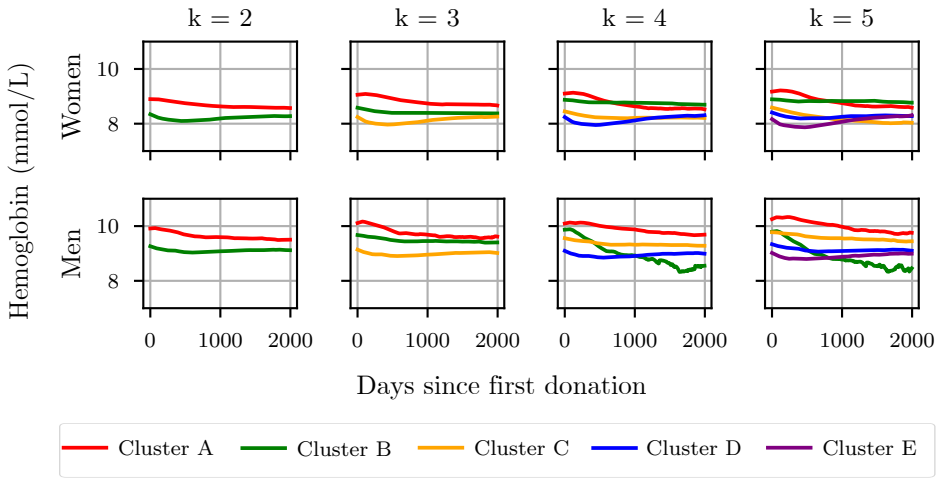

Figure 5.3: Cluster centroids after clustering resampled hemoglobin trajectories of 5000 female and 5000 male donors based on the intercept and slope of the linear trend, using the k-means clustering algorithm ($k \in [2, 3, 4, 5]$).

Sex	k	Cluster A $\bar{d}$ (N)	Cluster B $\bar{d}$ (N)	Cluster C $\bar{d}$ (N)	Cluster D $\bar{d}$ (N)	Cluster E $\bar{d}$ (N)	Sum $\bar{d}$
Women	2	8.1 (2028)	6.8 (2972)				36689
	3	10.1 (1075)	6.5 (1727)	9.6 (2198)			43195
	4	8.2 (602)	10.8 (1181)	6.7 (1362)	10.5 (1855)		46318
	5	6.2 (839)	14.1 (1016)	8.5 (881)	11.9 (1761)	8.8 (503)	52431
Men	2	11.0 (2843)	11.1 (2157)				55156
	3	9.5 (831)	6.5 (1924)	8.7 (2245)			40113
	4	10.8 (389)	15.3 (961)	7.0 (2104)	6.4 (1546)		43392
	5	6.0 (386)	6.3 (1071)	7.1 (1378)	7.9 (445)	8.6 (1720)	37234

Table 5.2: The mean distance from the centroid to the time series ($\bar{d}$) and the number of time series in each cluster (N) after trend clustering. Dynamic time warping is used as distance measure for evaluation. The rightmost column shows the sum of the within-cluster distances.

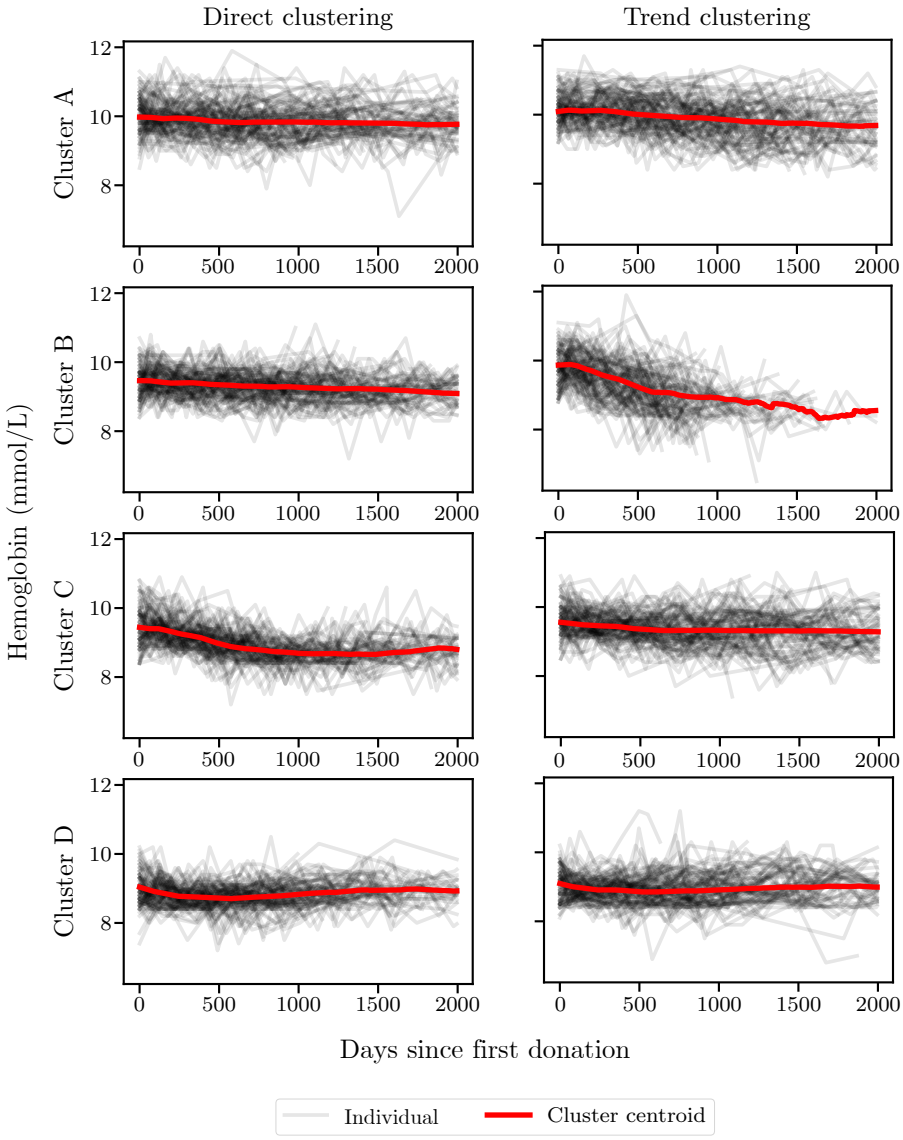

Figure 5.4: Cluster centroids after clustering resampled hemoglobin trajectories of 5000 male donors, using direct (left) or trend clustering (right) and k-means clustering with $k = 4$. Red lines are the cluster centroids. 100 randomly sampled individual hemoglobin trajectories from each cluster are also plotted to show the fit.

downward trend.

To verify the stability of the cluster centroids obtained by the k-means algorithm, we ran it several times with different random initialisation values. Visual inspection of the results showed that the algorithm consistently converged to the same centroids.

Discussion

The clusters obtained by the two methods are clearly very different. The centroids of the clusters are much more linear when the direct clustering is applied, compared to the trend feature clustering. The mean within-cluster distances are much smaller in the first method, which indicates denser clusters. However, this comparison is biased, because the first method used the same distance measure during clustering, so it is expected to minimise this distance. The second method minimised the Euclidean distance between the linear trend features of the time series, and not the DTW distances.

The results from the direct clustering are in line with our expectations. The clusters suit the time series relatively well, and the total sum of within-cluster distances decreases as the number of clusters increases. However, the time series are clustered together mostly on average hemoglobin level, which is not what we are looking for in this context. We would prefer to identify clusters based on the overall trends in each donor, so that we can distinguish donors with a high stable hemoglobin level, a low stable hemoglobin level, and decreasing hemoglobin levels from each other.

This is what we expected to see after clustering time series based on trend. It is partly what we see in the cluster centroids: in male donors, for $k = 4$ and $k = 5$, cluster B is very distinct from the others and has a steep downward slope. We know that declining hemoglobin trajectories are highly prevalent in female donors as well, but none of those centroids have a slope close to the one in male donors.

An interesting observation is that in almost all clusters, the hemoglobin level is decreasing in the first ± 500 days and then plateaus. This indicates that there is an initial effect of blood donation on average hemoglobin levels, but after the initial effect it reaches a new steady state. However, this is only based on the average hemoglobin levels of 5000 donors, and individual trajectories still show a lot of variation over time, making it hard to predict.

Limitations

There are some features in the data that were ignored in this first exploration in hemoglobin trajectory clustering. There is a seasonal component to hemoglobin levels: in warm seasons, levels are lower than in cold seasons. Because we used the number of days since first donation as time points and not the actual dates, we lost this information. An improvement would be to correct for seasonal variations before transforming the time variable. The same applies for the time of day hemoglobin was measured: it is highest in the morning and then drops steadily throughout the day.

A very clear feature of the data that was not used is the unequal sampling interval. Both methods required the intervals to be equal, so we resampled the time series using linear interpolation to satisfy this requirement. This means that we lose the information contained within the sampling intervals, and the resampled data points are of lower accuracy than the original measurements.

The third feature of the data that we would like to include in further analyses is whether or not a donation followed the hemoglobin measurement. If the hemoglobin level is below the threshold of 7.8 mmol/L for women or 8.4 mmol/L for men, no donation is made, and it is likely that the next measurement is higher. There is also an interaction with the interval length: if a donor has donated blood, the next measurement has to be at least 56 days later, but if the hemoglobin level was too low, it can be shorter.

Other Irregular Time Series Frameworks

There are many more fields in which irregular time series are observed (astronomy, medicine, economics, etc.), and in which the irregularities contain information we don't want to lose by transforming the data to equally spaced data. Some algorithms focus on calculating rolling time series operators such as simple moving averages or exponential moving averages. [68] This is a more elegant form of interpolation than what we have applied here, but the information contained in the intervals themselves is still lost.

A more fitting approach for our data might be a framework that takes two time series as input for each donor: one containing the hemoglobin measurements and one containing the interval lengths. We might consider a move to more complex algorithms, such as recurrent neural networks (RNNs) in combination with long short-term memory (LSTM) cells. [69] While the majority of RNN implementations still uses fixed time steps, the Phased LSTM model, which introduces an additional time

gate, handles irregular intervals without losing the information contained within the time steps. [70] A similar approach is Time-LSTM, which has been used to model website users' sequential actions by taking into account the sampling intervals. [71]

Another deep learning model that looks at informative missingness is GRU-D [72], which is based on gated recurrent units (GRU). It has been applied to real-world clinical data sets, where the missingness rate is highly correlated with variables of interest. This model has achieved good results in supervised classification tasks, and may also have useful applications for our unsupervised clustering task.

Future Work

By clustering donors' hemoglobin trajectories we hope to find clusters of donors that respond similarly to frequent blood donation. We assume that the clusters are a proxy for unobserved donor characteristics, such as iron intake, diet, physical activity levels and iron needs. If clustering is successful, we want to search for correlations between the cluster and donor information collected in questionnaires in previous studies carried out at Sanquin (Donor InSight). Eventually, the goal is to predict as early as possible in a donor's donation career which cluster they belong to, and to assign an optimal donation frequency based on this information. That way, deferral due to low hemoglobin may be minimised, and donors will stay healthy and motivated.