Received: 19 March 2020 | Revised: 25 May 2020 | Accepted: 3 June 2020

DOI: 10.1111/trf.15966

ORIGINAL RESEARCH

TRANSFUSION

Improved long-term time-series predictions of total blood use data from England

Anita K. Nandi[1] | David J. Roberts[2] | Asoke K. Nandi[3]

[1]Big Data Institute, University of Oxford, Oxford, UK

[2]Radcliffe Department of Medicine, National Health Service Blood and Transplant, Oxford Centre and BRC Haematology Theme, John Radcliffe Hospital, Oxford, UK

[3]Electronic and Computer Engineering, Brunel University London, Uxbridge, UK

Correspondence
Anita K. Nandi, Big Data Institute, University of Oxford, Oxford, OX3 7LF, UK.
Email: anita.k.nandi@gmail.com

Funding information
NHSBT R&D Funding and NIHR Program Grant "Erythropoiesis in Health and Disease" (NIHR-RP-PG-0310-1004)

Abstract

Background: Red blood cells are essential for modern medicine but managing their collection and supply to cope with fluctuating demands represents a major challenge. As deterministic models based on predicted population changes have been problematic, there remains a need for more precise and reliable prediction of use. Here, we develop three new time-series methods to predict red cell use 4 to 52 weeks ahead.

Study Design and Methods: From daily aggregates of red blood cell (RBC) units issued from 2005 to 2011 from the NHS Blood and Transplant, we generated a new set of non-overlapping weekly data by summing the daily data over 7 days and derived the average blood use per week over 4-week and 52-week periods. We used three new methods for linear prediction of blood use by computing the coefficients using Minimum Mean Squared Error (MMSE) algorithm.

Results: We optimized the time-window size, order of the prediction, and order of the polynomial fit for our data set. By exploiting the annual periodicity of the data, we achieved significant improvements in long-term predictions, as well as modest improvements in short-term predictions. The new methods predicted mean RBC use with a standard deviation of the percentage error of 2.5% for 4 weeks ahead and 3.4% for 52 weeks ahead.

Conclusion: This paradigm allows short- and long-term prediction of RBC use and could provide reliable and precise prediction up to 52 weeks ahead to improve the efficiency of blood services and sufficiency of blood supply with reduced costs.

1 | INTRODUCTION

Red blood cells are necessary for modern medicine in elective and emergency surgery, major trauma, hemorrhage, cancer care, and to support patients with congenital or acquired anemia.[1] The call-up of donors, scheduling of donor sessions, and manufacturing and supply of red blood cells to hospitals must be coordinated to match demand. Managing the collection and supply of red blood cells to cope with the fluctuating demand presents a major challenge for blood services. In spite of this, there are few tools available to accurately predict demand for either short-term or long-term planning. Any improvement of prediction tools would allow greater efficiency in the use of resources as well as a more resilient and secure blood supply chain.

Weekly use of red blood cells can change by 30% from week to week in our dataset and annual use can change by

3%-7% from year to year.[2,3] Predicting use for red blood cells in a simple deterministic model using the age-structure of the population, the age-specific incidence of disease, and the requirement of blood by indication and procedure for each disease has been attempted.[4-7] However, such models have consistently underestimated the changes in medical and transfusion practice.[8-10] Predictions made using projected population growth, number, and type of transfusion episodes overestimated demands.[5] There have been a wide variety of changes in medical and surgical management, such as introducing less invasive surgery and lowering the hemoglobin threshold for transfusion, which have made deterministic modeling highly prone to substantial errors.

An alternative strategy for prediction is to use time-series methods where any element in the time-series are assumed to be linearly related to previous elements by some mathematical relation with parameters that can be estimated. The estimated parameters can then be applied to extend the series into the future. The use of time-series methods for prediction have a long history.[11-15] There are a wide variety of time-series methods.[16] These approaches have been successfully applied in many fields including statistics,[17] communications,[18] signal processing,[19] adaptive noise cancellation,[20] earthquake prediction,[21] mathematical finance,[22] brain studies,[23,24] speech communication,[25] weather forecasting,[26] and econometrics.[27]

A previous study looked at time-series prediction of blood use,[28] and although these methods showed promising results, the deterioration of accuracy of predictions for long-term forecasting would limit long-term planning. In this paper we focus on the seasonality in the data with the aim of improving accuracy in long-term predictions of blood use. Seasonality is likely to be a significant factor due to strong seasonal patterns of activity in hospitals around variation in the number, type of admission, availability of capacity for elective surgery, and of staff.

Daily red blood cell use is readily available. Although red blood cell use varies significantly on a daily and weekly basis, in practice the window for useful predictions of future use are for 1 to 6 months to allow for matching of donor appointments and planning of donor sessions to predicted use. Predictions at longer intervals, such as a year ahead, are also useful to match the overall collection, manufacturing and issue capacity, and blood price to overall use, particularly as use falls.

Here we use three new time-series methods to predict red cell use 4 weeks to 52 weeks (1 year) ahead and demonstrate that the mean red cell use can be predicted with a standard deviation of the percentage error of 2.5% for 4 weeks ahead and 3.4% for 52 weeks ahead. By adjusting for recurring temporal and secular trends through a year including seasonal variation and holidays, significant improvements have been made from previous predictions, giving a standard deviation of the percentage error of 3.0% for 4 weeks ahead and 5.8% for 52 weeks ahead.[28] The proposed paradigm may form the basis for reliable short-term and long-term prediction of not only RBCs but also other components and even therapeutic procedures by blood services.

2 | MATERIALS AND METHODS

The focus of this paper lies in predicting the RBC use from 4 to 52 weeks ahead using a previously developed prediction paradigm,[28] but now incorporating three novel time-series methods. The three-stage prediction paradigm consists of: smoothing (eg, going from daily to monthly data) to reduce unhelpful noise; de-trending to extract and remove the long term variations from the data; and time-series modeling to accurately predict the remaining variations.

2.1 | Smoothing - data preparation

Daily aggregates of red blood cell units used cover a period of 6.5 years from February 1, 2005 to July 31, 2011 and were obtained from the NHS Department of Blood and Transport. We use aggregated data from seven consecutive days or integer multiples of seven consecutive days. This avoids both effects of daily variability as well as of variability between weekdays and weekends. A new set of non-overlapping weekly data was generated by summing the daily data over 7 days, ie, the first data

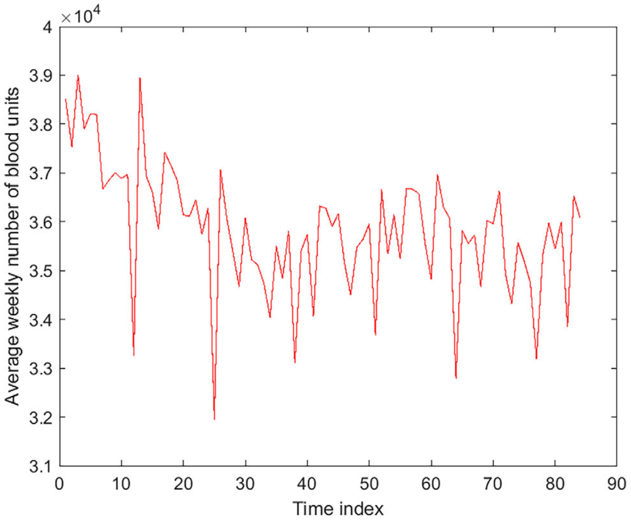

FIGURE 1 Average weekly blood use for each non-overlapping 4-week period from February 2005 to July 2011. This dataset contains 84 data points and are used for all prediction methods. The time index corresponds to the index of the 4-week period [Color figure can be viewed at wileyonlinelibrary.com]

point corresponds to the sum of Days 1-7, the second corresponds to Days 8-14, and so on; this new dataset of weekly blood use contains 338 data points. Most time-series methods used non-overlapping 4-week data, as shown in Figure 1. This was generated by summing the weekly data over 4 weeks and dividing by four, to give an average blood use per week over that 4-week period. In other words, the first data point is a weekly average blood use over Weeks 1-4, the second is a weekly average over Weeks 5-8, etc.; this non-overlapping 4-week dataset contains 84 data points.

Smoothed-overlapping data over a 52-week period, shown in Figure 2, was also used. This was generated by summing the weekly data over 52 weeks and dividing by 52, giving average blood use per week for that 52-week period, moving forward by 1 week each time. In other words, the first data point is a weekly average of Weeks 1-52, the second is a weekly average of Weeks 2-53, etc.; this overlapping 52-week dataset contains 287 data points. This generates a smoother time-series with less overall variation in average weekly number of blood units.

2.2 | Detrending

In the previous study, focusing on standard linear prediction, it was demonstrated that removing the underlying trend in the data, before applying time-series prediction methods, results in a very significant improvement in the accuracy of the prediction.[28] The trend is determined using a polynomial fit to the most recent w data points, where w is referred in this paper to as the time-window size.

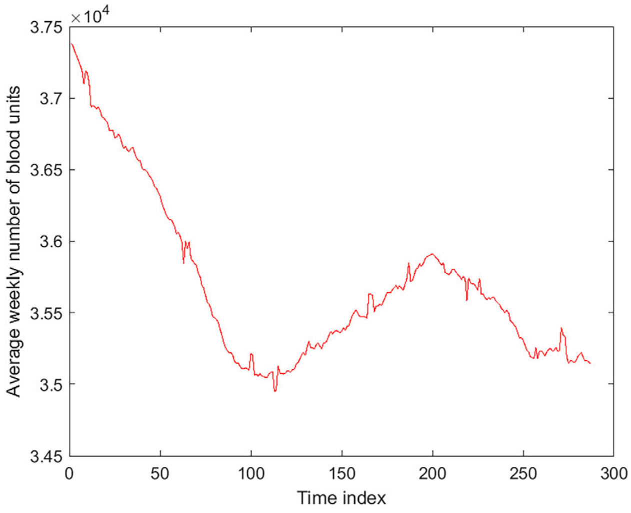

FIGURE 2 Average weekly blood use for overlapping 52-week periods, shifting by 1 week each time. This dataset contains 237 data points. The time index corresponds to the index of the overlapping 52-week period [Color figure can be viewed at wileyonlinelibrary.com]

Figure 3 shows a schematic of the steps taken to predict future blood use. It is interesting to note, as discussed later, for one of the methods that does not use standard linear prediction, it was found that removing the trend was not necessary for improving prediction and as such only the mean was removed.

2.3 | Time series methods

In this paper three new methods for predicting RBC use are explored that focus around Minimum Mean Squared Error (MMSE), aiming to improve the long-term prediction accuracy of blood use.

Time-series prediction methods use a set of previous data points in the time-series to predict future values. In general, it is assumed that the predicted value, $\hat{x}$, is some function of the past m values, as shown by,

$$\hat{x}(n+\alpha \mid n-1,n-2,\ldots,n-m)=f(x(n-1),x(n-2),\ldots,x(n-m)) \tag{1}$$

where n is the next time step in the series, α is the number of time steps ahead being predicted and x are the data points in the time-series. This defines m as the order of the prediction. In general, the function f is a non-linear function of the variables, but in this paper, we restrict the function f to be a linear function of the variables; this is known as linear prediction, which is illustrated by,

$$\hat{x}(n+\alpha)=\sum\nolimits_{i=1}^{m} a_i x(n-i) \tag{2}$$

where a_i are a set of coefficients to be estimated. The error in this linear prediction,

$e(n+\alpha)$, is defined to be,

$$e(n+\alpha)=x(n+\alpha)-\hat{x}(n+\alpha) \tag{3}$$

The linear time-series prediction problem lies in investigating methods for determining the a_i coefficients. There are several algorithms for linear prediction techniques, ie, methods for computing the coefficients a_i, that are well developed, eg, Minimum Mean Squared Error (MMSE) and Weighted Least Squares Error (WLSE).[12,16] However, there are circumstances when non-linear data analysis methods are required. Machine learning algorithms can be used to develop non-linear models for forecasting time-series data.[29-32] Examples of these algorithms include kernel-based machine learning, genetic programming, and artificial neural networks. Non-linear prediction methods are equally valid for the time-series data; however, they will not be considered in this paper.

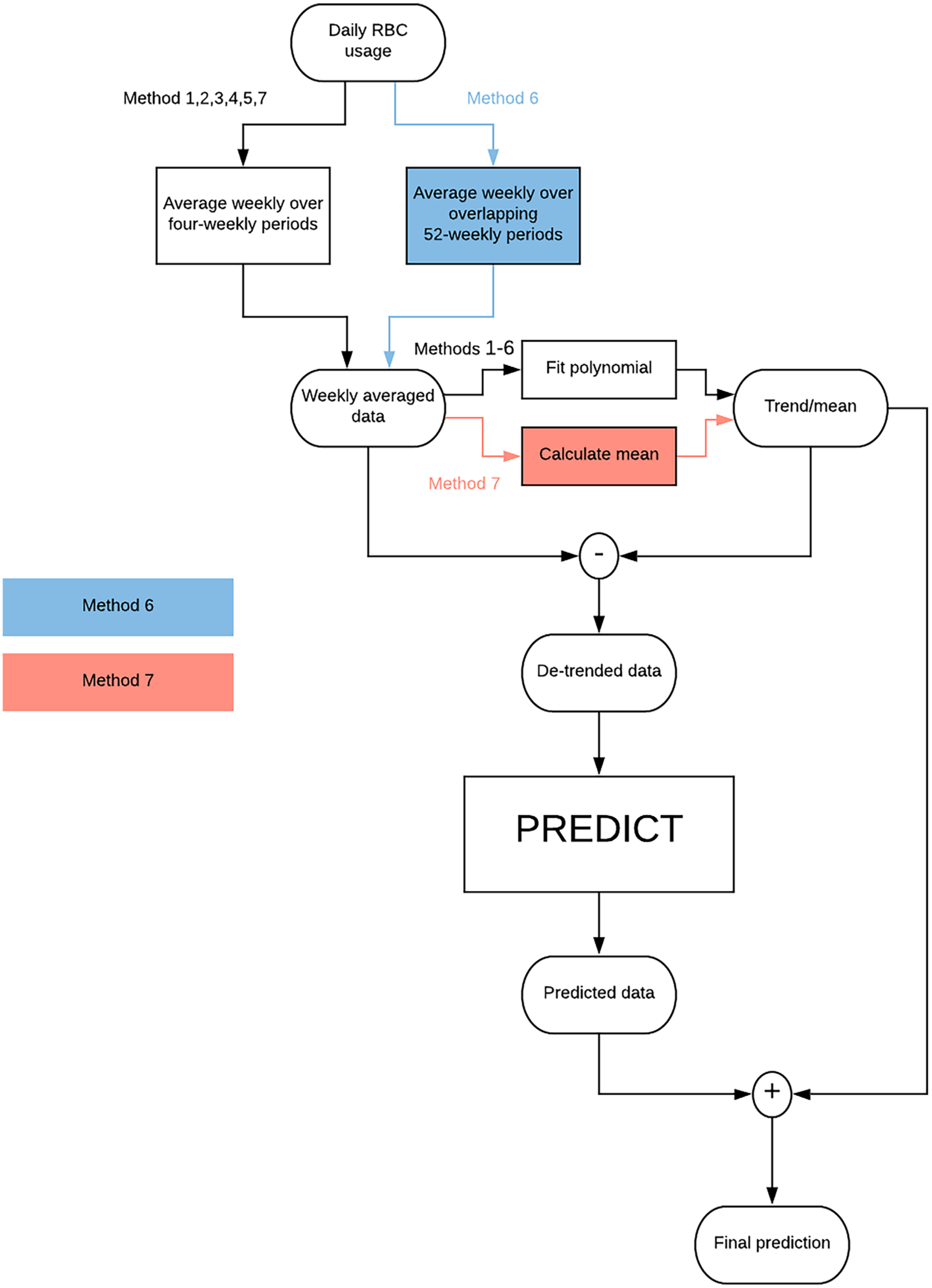

FIGURE 3 Schematic diagram of the processing steps involved in predicting future blood use. Rounded boxes represent data, while rectangles represent a processing stage. Variations to the processing for methods 6 (in blue) and 7 (in red) are shown in the diagram [Color figure can be viewed at wileyonlinelibrary.com]

First, MMSE provides an algorithm for determining the coefficients of the linear prediction based on minimizing the mean squared error, whose mathematical details are in Appendix A. This method is discussed in the previous study as Method 1.[28] Alternative methods based on the observation that the 4-week data contains some large dips and peaks, aimed to improve the prediction by mitigating the effect of these outliers. This can be achieved by using WLSE, different amounts of weighting account for the differences between Methods 2, 3, and 4 in the previous study.[28] Overall, there was not much variation in the predictions from these four methods.

Here, three new methods (Methods 5, 6, and 7) are developed with the aim of improving long-term prediction accuracy. As discussed, Method 1 uses standard MMSE, which computes the coefficients a_i from Equation (2) by minimizing the mean squared error. A new method, Method 5, was then considered, which involves flipping the time series data in the time-window over, so the most recent data point is at the beginning. Then the trend and mean were taken out before calculating the coefficients, which will be called b_i. The data being used for prediction is given by, $\{d(w), d(w-1), ..., d(1)\}$, where w is the time-window size. Standard MMSE prediction is applied to the beginning of the window, as shown by,

$$d(w+1-m)=\sum_{i=1}^{m} b_i d(w+1-m+i) \tag{4}$$

However, this predicts the data point $d(w+1-m)$, which is already known. The value to be predicted is $d(w+1)$, which can be found by rearranging Equation (4) to give,

$$\hat{d}(w+1)=\frac{1}{b_m}\left[d(w+1-m)-\sum_{i=1}^{m-1} b_i d(w+1-m+i)\right] \tag{5}$$

This alternative method of predicting the next data point in the time series is referred here as backward MMSE. In order to control the uncertainty in this prediction, as discussed in Appendix B, we fix b_m to some chosen value of order unity, referred to in this article as β, and use MMSE to calculate the remaining $(m-1)$ coefficients. Now, instead of using Equation (4), we use,

$$d(w+1-m)=\sum_{i=1}^{m-1}[b_i d(w+1-m+i)]+\beta x(w+1) \tag{6}$$

The mathematical detail for this method can be found in Appendix C.

Method 6 involves applying MMSE to the overlapping 52-week data described in Smoothing - data preparation. The 52-week dataset is smoother, making predictions easier, however after prediction the result must be transformed into a 4-week prediction. Each data point in this dataset contains 1 week of new information, therefore, in order to predict the next 4 weeks it is necessary to predict four data points ahead in the 52-week smoothed data, ie, $\alpha = 0$ for the 4-week data corresponds to $\alpha = 3$ for the 52-week smoothed data.

All the methods so far have used the most recent m data points, as shown in Equation (2). However, when predicting long-term blood use, the volume of blood issued for the same week of the year in previous years may contain more useful information. With that in mind, Method 7 uses non-standard linear prediction by applying MMSE on the original 4-week data, instead of using the most recent m data points; it uses the most recent m data points at the same time in previous years,

$$\hat{x}(n+\alpha)=\sum_{i=1}^{m} a_i x(n+\alpha-13i)+a_0 \tag{7}$$

To take advantage of the annual variation of the data, this method uses information of blood use at the same time in previous years as opposed to the most recent information that is available.

2.4 | Figure of merit

Implementing each of the three new time-series methods, described in De-trending, gives a set of predictions, $\hat{x}(n)$, for each of their corresponding known true data values, $x(n)$. The percentage error for each data point was calculated, $100(x(n)-\hat{x}(n))/x(n)$. To assess quantitatively the accuracy of the prediction methods, the mean and the standard deviation of these percentage errors were calculated. Given that the mean percentage error is sufficiently small, it is more important that the standard deviation of the percentage errors is as small as possible, ie, the error in predictions does not vary by a large amount. Additionally, it is important to consider what proportion of the time is the prediction within a reasonable region around the true value. For the final results we also quote the percentage of predictions that lie within the ±5% range of the true value. In this paper we compare our results from Methods 5, 6, and 7, to those using standard MMSE (Method 1).

3 | RESULTS

3.1 | Optimizing the parameters

The prediction paradigm, incorporating four time-series methods, contain various parameters that can be altered, which would affect the accuracy of the prediction. These parameters include the time-window size (w), the order

TABLE 1 Optimization of the prediction parameters: (A) fixed coefficient, β, for Method 5, (B) order of polynomial fit for Method 6, fixing w = 26 and m = 5, (C) order of polynomial fit for Method 7, fixing w = 26 and m = 5, (D) number of parameters for Method 7, fixing w = 26 and d = 0. Results of prediction applied to blood use data when predicting 4 weeks ahead ($\alpha = 0$), 8 weeks ahead ($\alpha = 1$), 12 weeks ahead ($\alpha = 2$), 16 weeks ahead ($\alpha = 3$), 20 weeks ahead ($\alpha = 4$), 24 weeks ahead ($\alpha = 5$), 28 weeks ahead ($\alpha = 6$), and 52 weeks ahead ($\alpha = 12$), for a range of parameter values are shown. In each box, corresponding to each experiment, the first number is the mean percentage error and the second number is the standard deviation of the percentage errors

(A)

	β					
α	**0.9**	**1**	**5**	**9**	**13**	**17**
0	0.38	0.37	0.27	0.26	0.26	0.26
	3.78	3.61	2.90	2.89	2.89	2.89
1	0.36	0.35	0.28	0.27	0.27	0.26
	4.72	4.47	3.16	3.09	3.07	3.06
2	0.33	0.32	0.28	0.28	0.28	0.28
	5.39	5.12	3.58	3.47	3.43	3.41
3	0.35	0.35	0.37	0.37	0.37	0.37
	4.70	4.47	3.35	3.31	3.30	3.30
4	0.43	0.42	0.34	0.33	0.33	0.33
	4.22	4.07	3.57	3.59	3.60	3.61
5	0.40	0.39	0.37	0.37	0.36	0.36
	4.83	4.67	3.98	3.97	3.96	3.96
6	0.38	0.38	0.34	0.34	0.34	0.34
	5.63	5.42	4.38	4.33	4.31	4.30
12	−0.08	−0.08	−0.07	−0.07	−0.07	−0.07
	7.01	6.83	6.05	6.04	6.03	6.03

(B)

	Order of the polynomial fit, *d*		
α	**1**	**2**	**3**
0	0.36	0.14	0.14
	3.13	3.15	3.95
1	0.26	−0.03	−0.04
	2.81	3.22	5.24
2	0.34	−0.14	−0.48
	2.80	3.81	8.09
3	0.24	0.02	−0.53
	2.75	3.87	10.2
4	0.44	0.26	−0.82
	2.84	4.16	13.9
5	0.40	0.42	−0.70
	3.04	4.25	17.8
6	0.56	0.58	0.20
	3.27	4.44	20.9
12	0.62	0.20	−1.12
	3.80	7.97	56.7

TABLE 1 (Continued)

(C)

	Order of the polynomial fit, *d*		
α	**0**	**1**	**2**
0	0.10	0.23	0.07
	2.49	2.65	2.63
1	0.10	0.22	0.06
	2.54	2.70	2.69
2	0.12	0.24	0.06
	2.62	2.79	3.04
3	0.22	0.35	0.18
	2.64	2.82	3.14
4	0.26	0.41	0.26
	2.81	3.06	3.52
5	0.32	0.48	0.29
	2.90	3.19	3.83
6	0.40	0.56	0.27
	2.98	3.27	3.99
12	0.75	0.97	−0.48
	3.42	3.94	5.71

(D)

	Number of parameters	
α	**2**	**3**
0	0.10	0.19
	2.49	2.68
1	0.10	0.22
	2.54	2.79
2	0.12	0.32
	2.62	2.83
3	0.22	0.36
	2.64	2.73
4	0.26	0.31
	2.82	2.88
5	0.32	0.23
	2.90	2.86
6	0.40	0.14

TABLE 1 (Continued)

(D)		
	Number of parameters	
α	2	3
	2.98	2.71
12	0.75	0.04
	3.42	3.60

of the prediction (m), and the order of the polynomial fit (d). For Method 5, there is an additional parameter of the fixed coefficient (β). An important advantage of this prediction paradigm is that parameters can be optimized for different situations. In the previous study, the optimal parameters for Methods 1-4 were found to be $w = 26$, m $= 5$, and $d = 2$.[28] As we are now using three new methods focusing on different aspects of the data, we must reconsider the optimal parameters for the new methods.

Method 5 uses the same data as Methods 1-4, so the same parameter values are used. However, as discussed in Section 2.3, Method 5 requires fixing the coefficient b_m to some chosen value, β, in order to control the prediction error, which we must optimize for the current dataset. Table 1A shows a significant improvement in the prediction performance as β increases to five, but for $\beta > 5$ the quality of the prediction starts to plateau, ie, there does not seem to be an upper limit on β. Based on this investigation Method 5 was carried out using a value of $\beta = 9$.

Before any of the prediction methods can be applied the trend in the data must be removed, as discussed in De-trending. As the data used in Methods 6 and 7 are different to previous methods, due to the data smoothing applied to leverage different aspects of the data, the order of the polynomial must be investigated for these new methods. Table 1b shows that a polynomial fit of $d = 1$ provides the best predictions for the 52-week smoothed data, as for large values of α the error in the polynomial fit is exaggerated, resulting in the second order polynomial fit giving much greater errors than a linear polynomial fit. As Method 7 does not use the standard linear prediction method given by Equation (2), it can no longer be assumed that removing the trend before applying the prediction improves the prediction method. Table 1c shows that the predictions made using Method 7 are significantly improved when applying the method without removing the trend first. Therefore, Method 6 was carried out using a polynomial fit with $d = 1$, and Method 7 was carried out using $d = 0$, ie, the trend was not removed before applying the prediction method, only the mean was subtracted.

TABLE 2 Optimized parameter values for each of the Methods 1-7: time-window size (w), order of prediction (m), order of polynomial fit (m), and fixed coefficient for backward MMSE (β). Methods 1-4 all use the same parameter values. The additional parameter β is only required for the backward MMSE prediction used in Method 5

Method	w	m	d	β
1-4	26	5	2	-
5	26	5	2	9
6	26	5	1	-
7	26	1	0	-

Due to the annual periodicity of the data, in Method 7 it may be beneficial to use a time-window size that is a multiple of 13 (corresponding to a year in 4-week data). According to Equation (7), the minimum window size is $13(m + 1)$. There are 84 data points in the 4-week data; as this is a small dataset, it would be beneficial to have a small a window size as possible to be able to predict more data points, ie, $m < 3$. Therefore, for Method 7 the window size could either be $w = 26$ or $w = 39$, ie, $m = 1$ or $m = 2$, respectively. A value of $m = 1$ corresponds to a two-parameter prediction and a value of $m = 2$ corresponds to a three-parameter prediction. Table 1d shows that using Method 7 with two parameters seems to be better for the prediction. Also, using two parameters corresponds to $w = 26$, which maintains consistency with the other methods and so allows for validity of comparisons. Therefore, Method 7 was carried out using two parameters ($m = 1$).

Final parameter values used for each of the Methods 1-7 are shown in Table 2.

3.2 | Comparison of the time-series methods

Each box in Table 3 shows the mean error, the standard deviation of the errors, as well as the percentage of predictions that lie within ±5% of the true value. These results are given for standard MMSE (Method 1) along with each of the three new prediction methods presented in this paper (Methods 5, 6, and 7). Predictions are made from one to seven 4-week periods ahead, as well as 1 year ahead (thirteen 4-week periods ahead), ie, 4-week, 8-week, 12-week, 16-week, 20-week, 24-week, 28-week, and 52-week. Plots of the predictions for 4 weeks ahead and 52 weeks (1 year) ahead are shown in Figures 4 and 5, respectively.

The total blood use data has been predicted for the next 4-week period with a standard deviation in the error

TABLE 3 Results for each of the standard MMSE and three new prediction methods applied to blood use data to predict 4 weeks ahead ($\alpha = 0$), 8 weeks ahead ($\alpha = 1$), 12 weeks ahead ($\alpha = 2$), 16 weeks ahead ($\alpha = 3$), 20 weeks ahead ($\alpha = 4$), 24 weeks ahead ($\alpha = 5$), 28 weeks ahead ($\alpha = 6$), and 52 weeks ahead ($\alpha = 12$). In each box, corresponding to each experiment, the first number is the mean percentage error, the second number is the standard deviation of the percentage errors, and the third number is the percentage of predictions that lie within ±5% of the true value

	Method			
α	1	5	6	7
0	0.28	0.26	0.36	0.10
	2.97	2.89	3.13	2.49
	95	95	88	95
1	0.08	0.27	0.26	0.10
	3.00	3.09	2.81	2.54
	95	95	95	95
2	0.07	0.28	0.34	0.12
	3.18	3.47	2.80	2.62
	89	88	93	95
3	0.06	0.37	0.24	0.22
	3.21	3.31	2.75	2.64
	89	89	95	95
4	0.19	0.33	0.44	0.26
	3.62	3.59	2.84	2.82
	83	87	94	93
5	0.19	0.36	0.40	0.32
	4.02	3.97	3.04	2.90
	85	85	92	92
6	0.14	0.34	0.56	0.40
	4.16	4.33	3.27	2.98
	81	77	88	92
12	−0.31	−0.07	0.62	0.75
	5.78	6.04	3.80	3.42
	59	57	76	85

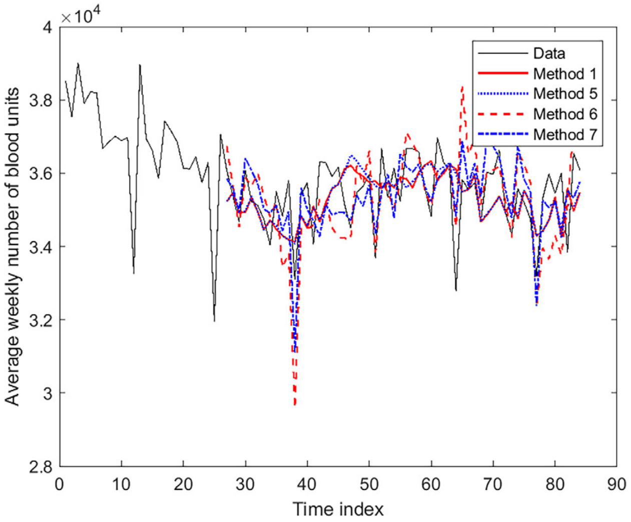

FIGURE 4 Results of the predictions using all four methods to predict the next 4-week period. The data is shown in red. The time index corresponds to the index of the 4-week period [Color figure can be viewed at wileyonlinelibrary.com]

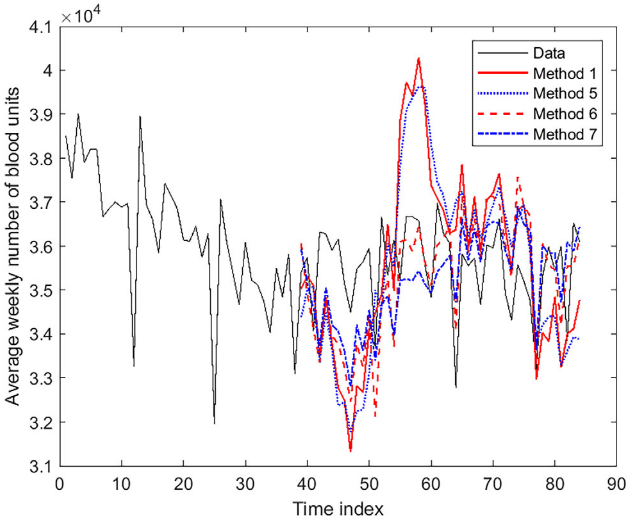

FIGURE 5 Results of the predictions using all four methods to predict a year ahead (12 four-week periods ahead). The data is shown in red. The time index corresponds to the index of the 4-week period [Color figure can be viewed at wileyonlinelibrary.com]

of 2.5%, with 95% of the predictions lying within 5%. The predictions for 52 weeks ahead achieve a standard deviation in the error of about 3.4%, with 85% of the predictions lying within 5% of the true value. The methods show similar performance for short-term predictions (1-6 months ahead), but Method 7 shows significantly improved performance when predicting more than 6 months ahead.

As there are seven different time-series methods in total, for each data point there exists seven different predictions. These can be combined by calculating the average of different prediction methods, but this was found to show no significant improvement to the results.

4 | DISCUSSION

Here we have evaluated our proposed prediction paradigm, incorporating three new time-series methods, to past RBC use data to make predictions 4 weeks, 8 weeks, 12 weeks, 16 weeks, 20 weeks, 24 weeks, 28 weeks, and 52 weeks ahead. These results show significant improvements on previous predictions of blood use using time-series data, especially for long-term predictions of more than 6 months ahead. As such the application of these methods may improve the effective planning of collection to the benefit of donors and blood services.

The standard MMSE prediction (Method 1) performs almost as well as other methods for short-term predictions, but the performance degrades significantly when predicting long-term use beyond 6 months ahead. However, the performance of Method 7 (using data points from the same time in previous years) remains impressive for up to a year ahead, with the potential to extend predicting further ahead. The ability to accurately predict long-term blood use is important for planning changes to the future blood collection strategy. Being prepared for changes to blood demand a year ahead presents the opportunity to effectively predict income and plan a more efficient use of resources throughout the blood supply chain.

The Method 7 provided predictions of aggregate use for 4 weeks ahead with a standard deviation of 2.5%, with 95% of the predictions lying within 5% of the true value, and for 52 weeks ahead with a standard deviation in of 3.4% with 85% of the predictions lying within 5% of the true value. For predicting 4 weeks ahead, of the 5% of predictions that lie outside 5% of the true value, a third overestimate use. The maximum surplus for any individual prediction was 2047 blood units, while the maximum deficit was 2048 blood units. For predicting 52 weeks ahead, of the 15% of predictions that lie outside 5% of the true value, 29% overestimate use (maximum surplus of 2260 units) and 71% underestimate use (maximum deficit of 2602 units).

These margins of error would be operationally acceptable as the current average weekly use of RBC units in England are approximately 27000 units or 3800 units per day averaged over 9 months. The current stock levels of red blood cells in the blood services and in hospitals are currently maintained at between 8- and 10-days' supply. Therefore, the blood supply chain could tolerate fluctuation in stock of 4000 units in any 1 week. In practice, adjustments to the supply could be made to cover such variation by minor changes to the collection schedule to maintain stable stock levels.

Previous attempts at predicting medium-term use for a group of patients or within a region or country have relied on simple linear extrapolation of year-on-year trend.[33,34] Generally, these methods have predicted a rising demand for blood based on demographics where the proportion of people older than 75 years is rising, eg, in North America and Europe. In turn, these models generated concern about potential shortfall in the supply of blood from younger donors.[35,36] However, these attempts for medium-term forecasting have been inaccurate and were unable to predict the trends in reduced blood use due to changes in medical and surgical practice as well as patient blood management.[37,9] These methods have been unsuccessful in accurately predicting medium or long-term trends, and short-term planning has relied on time-series methods from proprietary packages. This paper has developed time-series methods that produce accurate short-, medium-, and long-term predictions of blood use.

It is clear from this data set that seasonality was a significant factor in modulating use. This is likely to be due to strong seasonal patterns of activity in hospitals around variation in the number, type of admission, availability of capacity for elective surgery, and of staff. With this dataset it is not possible to establish the exact reasons for such seasonality but further exploration using hospital-level data may help analyze these trends and delineate causal factors in the seasonality of blood use.

Further improvements could be made if information were available on changes in surgical procedure or practices in transfusion medicine and how they are being implemented in the different regions. The data could also be examined by location or blood group to provide a more targeted call-up of donors, however the benefits of this may be offset by the increased random error in dealing with fewer blood units.

Also, there is certainly low but significant waste of whole blood at blood centers and in hospital transfusion services due to expiry of units beyond their mandated shelf-life. This study uses national data to look at overall blood use where the primary purpose is to allow better short- and medium-term matching of collection and demand. Waste of blood in hospitals would be addressed by different models using information based on regional or hospital level data, which is beyond the scope of this study.[38]

These findings of improved predictions, especially long-term predictions, using several time-series methods that are tailored to the specific data sets, potentially represent a significant advance in the techniques available to predict use. The improved predictions with reduced errors could allow greater efficiency in the call-up of donors, scheduling of donor sessions, and manufacturing and supply of RBCs to match demand.

In conclusion, it is important to appreciate that a straightforward use of time-series methods would not have produced as good results as presented in this paper. By exploiting the annual periodicity of the time-series, we were able to improve significantly long-term predictions of blood use, with anticipated commensurate improvement in the effectiveness and efficiency of collection. These methods are in principle capable of further improvements using more granular local data and by more precise alignment of the methods with the data.

ACKNOWLEDGMENTS

We would like to acknowledge the help of National Health Service Department of Blood and Transplant (NHSBT) staff in supplying the data used in these analyses. David Roberts and Anita Nandi were in part supported by NHSBT R&D Funding; also, David Roberts, Anita Nandi, and Asoke Nandi were in part supported by the NIHR Program Grant "Erythropoiesis in Health and Disease" (NIHR-RP-PG-0310-1004).

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

ORCID

Anita K. Nandi https://orcid.org/0000-0002-5087-2494
Asoke K. Nandi https://orcid.org/0000-0001-6248-2875

REFERENCES

1. Klein H (2014). Red cell transfusion. In: Klein H, Anstee D, editors. Mollison's Blood Transfusion In Clinical Practice. 12th ed. Oxford, UK: Wiley-Blackwell.
2. Williamson LM, Devine DV. Challenges in the management of the blood supply. Lancet 2013;381:1866-75.
3. Tinegate H, Chattree S, Iqbal A, et al. Ten-year pattern of red blood cell use in the North of England. Transfusion 2013;53:483-9.
4. Wells AW, Mounter PJ, Chapman CE, et al. Where does blood go? Prospective observational study of red cell transfusion in north England. BMJ 2002;325:803-6.
5. Greinacher A, Fendrich K, Alpen U, et al. Impact of demographic changes on the blood supply: Mecklenburg-West Pomerania as a model region for Europe. Transfusion 2007;47:395-401.
6. Greinacher A, Fendrich K, Brzenska R, et al. Implications of demographics on future blood supply: a population-based cross-sectional study. Transfusion 2011;51:702-9.
7. Drackley A, Newbold KB, Paez A, et al. Forecasting Ontario's blood supply and demand. Transfusion 2012;52:366-74.
8. Tinegate H, Pendry K, Murphy M, et al. Where do all the red blood cells (RBCs) go? Results of a survey of RBC use in England and North Wales in 2014. Transfusion 2016;56:139-45.
9. Greinacher A, Weitmann K, Lebsa A, et al. A population-based longitudinal study on the implications of demographics on future blood supply. Transfusion 2016;56:2986-94.
10. Greinacher A, Weitmann K, Schönborn L, et al. A population-based longitudinal study on the implication of demographic changes on blood donation and transfusion demand. Blood Adv 1975;1:867-74.
11. Wiener N. Extrapolation, Interpolation, and Smoothing of Stationary Time Series, Cambridge, MA: MIT Press; 1949.
12. Mandel J. The Statistical Analysis of Experimental Data. New York: Interscience; 1964.
13. Makhoul J. Linear prediction: a tutorial review. Proc IEEE 1975;63:561-80.
14. Hamilton J. Time Series Analysis, Princeton, NJ: Princeton University Press; 1994.
15. Hayes MH. Statistical Digital Signal Processing and Modeling. New York: John Wiley & Sons; 1996.
16. Haykin SO. Adaptive Filter Theory. 5th ed. Upper Saddle River, New Jersey: Prentice Hall; 2013.
17. Falk M, Marohn F, Michel R, et al. A first course on time series analysis: examples with SAS, GNU Free Documentation License. https://www.uni-wuerzburg.de/fileadmin/10040800/user_upload/time_series/the_book/2011-March-01-times.pdf. Accessed July 11 2020.
18. Rappaport TS. Wireless Communications: Principles and Practice, Vol. 2. Prentice Hall PTR: Upper Saddle River, New Jersey; 1996.
19. Carter GC. Coherence and time delay estimation. Proc IEEE 1987;75:236-55.
20. Zarzoso V, Nandi AK. Non-invasive fetal electrocardiogram extraction: blind separation versus adaptive noise cancellation. IEEE Trans Biomed Eng 2001;48:12-8.
21. Varotsos P, Sarlis NV, Skordas ES. Natural Time Analysis: The New View of Time: Precursory Seismic Electric Signals, Earthquakes and Other Complex Time Series. Secaucus, New Jersey: Springer Science & Business Media; 2011.
22. Tsay RS. Financial Time Series. In: Wiley StatsRef: Statistics Reference Online. New Jersey: John Wiley and Sons; 2014. p. 1-23.
23. Goutte C, Toft P, Rostrup E, et al. On clustering fMRI time series. Neuroimage 1999;9:298-310.
24. Mormann F, Andrzejak RG, Elger CE, et al. Seizure prediction: the long and winding road. Brain 2006;130:314-33.
25. Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans Acoust Speech Signal Process 1978;26:43-9.
26. Craddock JM. The analysis of meteorological time series for use in forecasting. J R Stat Soc Ser D Stat 1965;15:167-90.
27. Enders W. Applied Econometric Time Series. 4th ed.Hoboken, NJ: John Wiley & Sons; 2015.
28. Nandi AK, Roberts DJ, Nandi AK. Prediction paradigm involving time-series applied to total blood issues data from England. Transfusion 2020;60:535-43.
29. Franses PH, Draisma G. Recognizing changing seasonal patterns using artificial neural networks. J Econom 1997;81:273-80.
30. Elkateb MM, Solaiman K, Al-Turki Y. A comparative study of medium- weather-dependent load forecasting using enhanced artificial/fuzzy neural network and statistical techniques. Neurocomputing 1998;23:3-13.
31. BuHamra S, Smaoui N, Gabr M, et al. The Box–Jenkins analysis and neural networks: prediction and time series modelling. App Math Model 2003;27:805-15.
32. Zhang GP, Qi M. Neural network forecasting for seasonal and trend time series. Eur J Oper Res 2005;160:501-14.
33. Lau EHY, He X-Q, Lee C-K, et al. Predicting future blood demand from thalassemia major patients in Hong Kong. PLoS One 2013;8:e81846.
34. Borkent-Raven BA, Janssen MP, Van Der Poel CL. Demographic changes and predicting blood supply and demand in the Netherlands. Transfusion 2010;50:2455-60.
35. Weidmann C, Schneider S, Litaker D, et al. A spatial regression analysis of German community characteristics associated with voluntary non-remunerated blood donor rates. Vox Sang 2012;102:47-54.

36. Seifried E, Klueter H, Weidmann C, et al. How much blood is needed? Vox Sang 2010;100:10-21.

37. Greinacher A, Fendrich K, Hoffmann W. Demographic changes: the impact for safe blood supply. Transfus Med Hemother 2010;37:141-8.

38. Yu X, Wang Z, Shen Y, et al. Population-based projections of blood supply and demand, China, 2017-2036. Bull World Health Organ 2020;98:10-8.

How to cite this article: Nandi AK, Roberts DJ, Nandi AK. Improved long-term time-series predictions of total blood use data from England. *Transfusion*. 2020;1–12. https://doi.org/10.1111/trf.15966

APPENDIX A

Appendix A: derivation of the Weiner-Hopf equations

We want to predict the $(n + \alpha)$ data point using linear prediction,

$$\hat{x}(n+\alpha \mid n-1, n-2, \ldots, n-m) = \sum_{i=1}^{m} \mathrm{a_i} x(n-i)$$

The error in the prediction is given by,

$$e(n+\alpha) = x(n+\alpha) - \hat{x}(n+\alpha)$$

To apply MMSE we want to minimize the mean squared error, MSE,

$$MSE, f = E\left\{e(n)^2\right\}$$

$$= E\left\{[x(n+\alpha) - \hat{x}(n+\alpha)]^2\right\}$$

Therefore, we need to solve the equations,

$$\frac{\partial f}{\partial a_j} = 0 \,\text{for}\, j = 1, 2, \ldots, m$$

The left hand side is given by,

$$\frac{\partial f}{\partial a_j} = E\left\{\frac{\partial f}{\partial a_j}(x-\hat{x})^2\right\}$$

$$= E\left\{2(x-\hat{x})\frac{\partial}{\partial a_j}(x-\hat{x})\right\}$$

Using the fact that,

$$\frac{\partial \hat{x}}{\partial a_j} = x(n-j)$$

We end up with,

$$\frac{\partial f}{\partial a_j} = E\{2(x-\hat{x})x(n-j)\} = 0$$

$$E\{\hat{x}(n+\alpha)x(n-j) - x(n+\alpha)x(n-j)\} = 0$$

$$E\left\{\left[\sum_{i=1}^{m} a_i x(n-i)\right] x(n-j)\right\} = E\{x(n+\alpha)x(n-j)\}$$

$$\sum_{i=1}^{m} a_i E\{x(n-i)x(n-j)\} = r_{xx}(\alpha+j)$$

Finally we get,

$$\sum_{i=1}^{m} a_i r_{xx}(\ j-i) = r_{xx}(\alpha+j) \ \ for j = 1, 2, \ldots, m$$

These are the Weiner-Hopf equations, which can be also written as a matrix equation,

$$\begin{bmatrix} r_{xx}(0) & r_{xx}(-1) & r_{xx}(-2) & \ldots & r_{xx}(1-m) \\ r_{xx}(1) & r_{xx}(0) & r_{xx}(-1) & \ldots & r_{xx}(2-m) \\ r_{xx}(2) & r_{xx}(1) & r_{xx}(0) & \ldots & r_{xx}(3-m) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{xx}(m-1) & r_{xx}(m-2) & r_{xx}(m-3) & \ldots & r_{xx}(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \ldots \\ a_m \end{bmatrix} = \begin{bmatrix} r_{xx}(\alpha+1) \\ r_{xx}(\alpha+2) \\ r_{xx}(\alpha+3) \\ \ldots \\ r_{xx}(\alpha+m) \end{bmatrix}$$

Appendix B: error analysis for backward MMSE

Define $\{d(n)\}$ as the data with the trend taken out. In standard MMSE we use the data points $\{d(1), d(2), d(3), d(4), d(5)\}$, assuming $m = 5$, to predict the next point, $\hat{d}_f(6)$,

$$\hat{d}_f(6) = a_1 d(5) + a_2 d(4) + a_3 d(3) + a_4 d(2) + a_5 d(1)$$

In Method 5, also called backward MMSE, we used the data points {$d(1)$, $d(2)$, $d(3)$, $d(4)$, $d(5)$} to predict the point $\hat{d}_b(6)$,

$$d(1) = b_1 d(2) + b_2 d(3) + b_3 d(4) + b_4 d(5) + b_5 \hat{d}_b(6)$$

$$\hat{d}_b(6) = \frac{1}{b_5}[d(1) - [b_1 d(2) + b_2 d(3) + b_3 d(4) + b_4 d(5)]]$$

We will assume, $\delta a_1 = \delta a_2 = \delta a_3 = \delta a_4 = \delta a_5 \equiv \delta a$, and equivalently, $\delta b_1 = \delta b_2 = \delta b_3 = \delta b_4 = \delta b_5 \equiv \delta b$. Consider the errors in the predictions for standard MMSE and backward MMSE,

$$\delta \hat{d}_f(6) = [d(1) + d(2) + d(3) + d(4) + d(5)]\delta a$$

$$\delta \hat{d}_b(6) = \frac{\delta b}{b_5}\left[d(2) + d(3) + d(4) + d(5) + \hat{d}_b(6)\right]$$

This shows that the error in the backward MMSE prediction is roughly scaled by a factor of $1/b_5$. Therefore, for b_5 less than unity, this will negatively affect the errors on the backward MMSE predictions. We address this problem by fixing the value of b_5, referred to as β, to be larger than unity. This parameter can be optimized for different datasets, but in this paper we use $\beta = 9$.

Appendix C: Weiner-Hopf equations for fixed b_m

We want to predict the $(n + \alpha)$ data point using linear prediction, but with the b_m coefficient fixed at β,

$$\hat{x}(n+\alpha | n-1, n-2, \ldots, n-m) = \sum_{i=1}^{m-1} \mathrm{b}_i x(n-i) + \beta x(n-m)$$

Following the same procedure as in Appendix A, we minimize the mean squared error by solving the equations,

$$\frac{\partial}{\partial b_j}\left[E\left\{[x(n+\alpha) - \hat{x}(n+\alpha)]^2\right\}\right] = 0 \text{ for } j = 1, 2, \ldots, m$$

Which gives the same results are before,

$$E\{\hat{x}(n+\alpha)x(n-j) - x(n+\alpha)x(n-j)\} = 0$$

We can now substitute in the expression for $\hat{x}$ to get the required equations.

$$E\left\{\left[\sum_{i=1}^{m-1} b_i x(n-i) + \beta x(n-m)\right] x(n-j)\right\} = E\{x(n+\alpha)x(n-j)\}$$

$$\sum_{i=1}^{m} a_i E\{x(n-i)x(n-j)\} + \beta E\{x(n-m)x(n-j)\} = r_{xx}(\alpha + j)$$

Finally we get,

$$\sum_{i=1}^{m-1} b_i r_{xx}(j-i) + \beta r_{xx}(m-j) = r_{xx}(\alpha + j) \text{ for } j = 1, 2, \ldots, m$$

These are the modified Weiner-Hopf equations, which can also be written in matrix form,

$$\begin{bmatrix} r_{xx}(0) & r_{xx}(-1) & r_{xx}(-2) & \ldots & r_{xx}(2-m) \\ r_{xx}(1) & r_{xx}(0) & r_{xx}(-1) & \ldots & r_{xx}(3-m) \\ r_{xx}(2) & r_{xx}(1) & r_{xx}(0) & \ldots & r_{xx}(4-m) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{xx}(m-2) & r_{xx}(m-3) & r_{xx}(m-4) & \ldots & r_{xx}(0) \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \ldots \\ b_m \end{bmatrix} = \begin{bmatrix} r_{xx}(\alpha+1) - \beta r_{xx}(m-1) \\ r_{xx}(\alpha+2) - \beta r_{xx}(m-2) \\ r_{xx}(\alpha+3) - \beta r_{xx}(m-3) \\ \ldots \\ r_{xx}(\alpha+m) - \beta r_{xx}(1) \end{bmatrix}$$