IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 19, 2017, accepted March 7, 2017, date of publication May 31, 2017,
date of current version December 22, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2696502

Trend Analysis of Fragmented Time Series for mHealth Apps: Hypothesis Testing Based Adaptive Spline Filtering Method With Importance Weighting

XIANGFENG DAI, (Member, IEEE), AND MARWAN BIKDASH, (Member, IEEE)
Department of Computational Science and Engineering, North Carolina A&T State University, Greensboro, NC 27401, USA
Corresponding author: Marwan Bikdash (bikdash@ncat.edu)

ABSTRACT The growth of mobile devices has provided significant opportunities for developing healthcare apps based on the mobile device ability to collect data. Unfortunately, the data collection is often intermittent. Missing data present significant challenges to trend analysis of time series. Straightforward approaches consisting of supplementing missing data with constant or zero values or with linear trends can severely degrade the quality of the trend analysis. In this paper, we present a robust adaptive approach to discover the trends from fragmented time series. The approach proposed in this paper is based on the hypothesis-testing-based adaptive spline filtering (HASF) trend analysis algorithm, which can accommodate non-uniform sampling and is therefore inherently robust to missing data. HASF adapts the nodes of the spline based on hypothesis testing and variance minimization, which adds to its robustness. Further improvement is obtained by filling gaps by data estimated in an earlier trend analysis, provided by HASF itself. Three variants for filling the gaps of missing data are considered, the best of which seems to consist of filling significantly large gaps with linear splines matched for continuity and smoothness with cubic splines covering data-dense regions. Small gaps are ignored and addressed by the underlying cubic spline fitting. Finally, the existing measurements are weighted according to their importance by simply transferring the importance of the missing data to their existing neighbors. The methods are illustrated and evaluated using heart rate data sets, blood pressure data sets, and noisy sine data sets.

INDEX TERMS Trend analysis, missing measurements, mobile health, mHealth, adaptive filtering, hypothesis testing, fragmented time series, health behavior change, missing observation, smartphone, digital health, health care.

I. INTRODUCTION

mHealth, also known as mobile health, denotes the use of mobile devices in medicine, healthcare and public health services [40]. The use of mHealth apps is rapidly increasing and expected to continue to grow in the coming years. The total market size of mobile health is predicted to reach 60 billion dollars by 2020 [54]. mHealth apps contribute to the exponential increase in digital interventions for educating users about preventive health care services, supporting weight loss, helping users manage their energy balance, promoting healthy behavior change, and assisting chronic disease management [1]. mHealth apps have been also used for disease surveillance, treatment support and epidemic outbreak tracking. They have extraordinary potential and have demonstrated significant outcomes across different types of populations [24], [25] since they combine the benefits of e-Health [13], [15] with the ubiquity of mobile devices.

Trend analysis plays an important role in mHealth apps, such as sleep tracking [16], [21], [50], [51], anxiety and mood tracking [37], weight tracking [19], [29], heart rate tracking [3] and workouts tracking [16], [17], [19]–[21], [43], [55]. However, it is quite common for mHealth apps to have missing data of various lengths [41] as shown in Table 1. For example, a sleep tracking app will have missing data if the mobile device has trouble detecting movements in bed; a workout tracking app will have missing data if the battery is down, or users forget to input moods on their mood tracking apps. Gaps in time series can present significant challenges

TABLE 1. Existing mHealth Apps.

mHealth Apps	Smoothing
Beddit Sleep Tracker [3]	Yes
Fitbit [16]	Yes
FitPort [17]	Yes
HealtheConnect [19]	No
Health Mate [20]	Yes
Jawbone UP [21]	Yes
Lose It! [29]	Yes
Pacifica [37]	No
Pillow [42]	No
Record by Under Armour [43]	No
Sleep Cycle Alarm Clock [50]	Yes
SleepBot [51]	No
StepsApp Pedometer [55]	Yes

to time series analysis, especially if the gaps are wide and multiple.

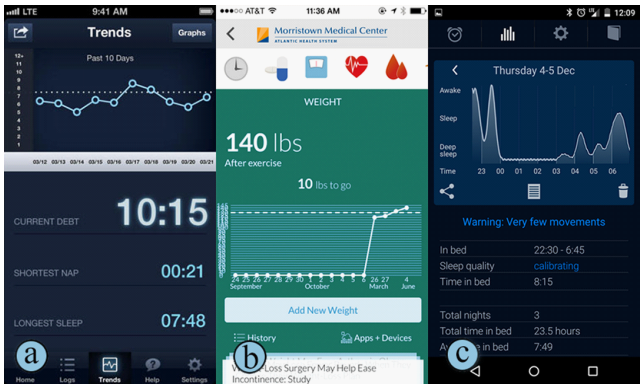

FIGURE 1. Examples of Trend Analysis of Existing mHealth Apps: (a) no smoothing [51], (b) no smoothing [19], (c) smoothing [50], (b and c) the holes of missing data are represented as "0"s.

Unfortunately, the current mHealth apps often lack suitable methods for handling fragmented time series. For example, apps like HealtheConnect [19], Pacifica [37], Pillow [42], Record by Under Armour [43], and SleepBot [51] represent trends by simply connecting all data points without using a smoothing process. An illustration of this from the SleepBot app is shown in Figure 1 (a). Other apps such as Beddit Sleep Tracker [3], Fitbit [16], FitPort [17], Health Mate [20], Jawbone UP [21], Lose It! [29], Sleep Cycle Alarm Clock [50], and StepsApp Pedometer [55] use smoothed time series to expose the trends. However, these apps handle missing data by filling in the gaps with zeros, which significantly reduces the reliability of the trend analysis. Examples of this are demonstrated by the HealthConnect [19] app in Figure 1(b) and the Sleep Cycle [50] app in Figure 1(c).

The situation is further complicated by the existence of noise in the measurements. Unfortunately, the noise is often nonstationary, which adds to the complexity. The effect of noise is alleviated by standard filtering or smoothing algorithms, but none of the mHealth apps surveyed in this paper address data gaps beyond simply filling in the gaps with zeros or with straight lines that connect the data points at

FIGURE 2. Missing data are represented by straight lines [51].

the edges of the missing segments. The effect of representing missing data with constant or straight line interpolations on the statistics of the noise and the smoothed signals is not well-understood.

In this paper, we present three variant methods to fill gaps containing missing data in mHealth apps in a more consistent way. The Hypothesis Testing Based Adaptive Spline Filtering (HASF) method [10] is subsequently applied to the filled time series to compute the trends. The remainder of this paper is structured as follows. In Section II, we review the literature of handling missing data in time series. In Section III, three variants to fill gaps of missing data are proposed, and the importance-weighting method is introduced. Numerical evaluation is presented in Section IV. Section V concludes the paper.

II. RELATED WORK

Trend analysis in time series is widely used in different domains such as finance [5], [32], climate analysis and forecasting [47], [56], network security [26], public health [11], [12], [39] and biomedicine [22]. The missing data problem in fragmented time series is a common challenge in trend analysis. Most classical methods, e.g. those used in [18] and [33], make many assumptions about the noise statistics, such as independence and normality, and replacing missing data with constants or straight line values will violate those assumptions [2]. Other approaches use different filtering techniques to substitute plausible values for the missing observations. Some of these approaches include:

A. AR FILTERING

The standard and most basic filtering technique is to think of the information signal as being a very low-frequency signal which is corrupted by a wide-frequency but stationary noise. In this context, obtaining the information signal is an exercise of signal filter design. Moreover, most filtering methods assume uniform sampling, an assumption that missing data also violate. Anava *et al.* [2] sought to improve the performance of a predictor dealing with missing data by adapting the coefficients of an autoregressive (AR) predictor model until it appears to converge to the best

AR predictor *in hindsight*. In other words, earlier versions of an AR predictor are used to fill in the missing data, and newer AR predictors are then derived. Unfortunately, the method seems to assume stationarity in the statistics, and hence may not work well in situations where nonstationary changes in the trend are of primary importance as is the case in most mHealth applications. Mirsaidi *et al.* [31] also presented an approach to estimate the missing values based on AR model reconstruction where the available data remains undisturbed. This approach reconstructs the values well at the front side of the gap; however, the backside of the gap is not smooth and contains jumps in the reconstructed signal. Broersen *et al.* [8] proposed an improvement based on AR modeling, and the resulting likelihood maximization algorithm, ARFIL, was shown to perform well, even with 90% missing data. Their main emphasis, however, was in estimating spectra (and hence assumed stationarity). The performance of the method in detecting shifts in trends in nonstationary data is therefore not guaranteed.

B. BOX-JENKINS FILTERING

Other stationary methods have been proposed to estimate the missing data by using the filtering process. For instance, Lewellen and Vessey [28] used the well-known Box-Jenkins procedure [7] to smooth a fragmented genetic sequence and approximate the missing values using a linear interpolation. Damsleth [14] developed an optimal method that is based on a linear combination of the forward and backward forecasts for reconstructing the missing values using an ARMA model. Mahir and Al-Khazaleh [30] employed a similar approach using an ARIMA (autoregressive integrated moving average) model to reconstruct the values in the missing data segments.

C. VARIANCE CONSTRAINED FILTERING

Wang *et al.* [58] considered missing measurements subject to norm-bounded parameter uncertainties and predicted the missing values using a Variance Constrained Filter (VCF). Moreover, [59] designed a robust Finite-Horizon Filter (FHF) where the missing measurements are described by a binary switching sequence satisfying a conditional probability distribution. An upper bound for the state estimation error variance is first derived for all possible missing observations and all admissible parameter uncertainties.

D. KALMAN FILTERING

The Kalman filter and its variations are also widely used for missing measurements [27]. If the dynamics of the information signal are known, recursive models can be employed to model the series and estimate its model parameters [48]. Sinopoli *et al.* [49] addressed the missing data problem by checking the linear minimum mean square error (LMMSE) of the Kalman filter when the missing data appears. Similar studies such as [9], [35], [44], and [52] used the Kalman filter, where the missing data are modeled using a jump linear system (JLS). In most of these papers, one must have a reasonable model for the process generating the time series or be willing to estimate or refine the parameters of the generating process, which makes the estimation problem nonlinear and complicated and necessitates the use of Extended Kalman filtering techniques.

E. OTHER APPROACHES

Scargle [45] studied astronomical time series and reconstructed the missing data by computing Fourier coefficients as the least squares fit of sines and cosines to the available remaining observations. However, this approach is not quite stable when describing slopes and background shapes in the spectrum [6], [8]. Nichols [34] proposed a complex three-stage procedure to estimate the missing data by smoothing the data, estimating characteristics of the signal and its spectrum, and then jointly estimating the missing data and the characteristics using nonlinear optimization.

Owili *et al.* [36] employed an optimal linear interpolation approach where the missing values in a time series are generated by a bilinear model. However, this approach can only be recommended in pure bilinear time series data whose innovations have a student-t distribution.

In the context of Doppler data, another approach [57] is not to assume any model generating the time series and to simply interpolate the samples using straight lines or splines, but the authors in [57] argue that it does not produce consistent results. They classified such methods into two categories: simple (largely based on resampling an irregularly-sampled underlying signal) and complex. They noted that linear interpolation is a more robust resampling method but that the estimated variance is too low. They provided a theoretical analysis of simple interpolation methods based on the assumption that the resampling period is much larger than the smallest original sampling period, perhaps leading to a loss of resolution.

Sehgal *et al.* [46] proposed a method by which missing DNA fragments in a gene can be reconstructed by comparing the damaged gene sequence with other similar gene sequences. Obviously, this method requires an understanding of the model generating the sequence, and may have application in mHealth. For instance, a heart rate sequence with missing data can be analyzed by comparing it with other intact sequences from the same patient or other similar patients. This approach requires more expert knowledge, such as a database of similar cases, and may be useful in deriving a *priori* estimates that can guide the estimation or filling-in of missing data.

F. NONSTATIONARITY ANALYSIS AND HASF

Bondon and Bahamonde [4] accounted for nonstationarity in the variance (heteroscedasticity). It was noted that the problem of estimating an autoregressive conditionally heteroscedastic model in the presence of missing values had not yet been addressed before. Soubeyrand *et al.* [53] studied fragmented nonstationary data in the context of determining whether two models governing climatological data behaved differently before and after a given date (presumably reflect-

ing major human effect). They used hypothesis testing to answer this question, but the data was uniformly sampled. In our previous work, we proposed a Hypothesis Testing Based Adaptive Spline Filtering (HASF) method [10]. The HASF method filters the data into flexible length sections and solves the challenging nonstationary time-series problem. Unlike other filtering methods, the HASF combines concepts from nonstationary time-series analysis, spline fitting and hypothesis testing.

III. HYPOTHESIS TESTING-BASED ADAPTIVE SPLINE FILTERING WITH MISSING DATA

In this study, we propose a two-step procedure to find the trends of fragmented time series. The first step is to substitute plausible values for the missing observations to initialize a complete time series. We present three variants: 1) fill gaps with straight lines, 2) fill gaps with cubic splines, and 3) fill gaps with cubic and linear filtering. The second step is to carry out the trend using the initialized filled time series using the HASF method [10]. In Section III-A, we review the HASF methodology developed in [10] at some length, as to later clarify the modifications discussed in Section III-B. The proposed method can be thought of as a special case of a more general iterative procedure where gaps are filled based on earlier trend analysis.

A. REVIEW OF HASF [10]

The basic concept of HASF is to break the time series into flexible sections, each of which is curve-fitted with a cubic spline, and to impose appropriate constraints such as continuity and smoothness between the sections, a minimum or maximum section length, etc. The number of sections and the nodes between them are adapted from the data using hypothesis testing applied to the second statistics of the residual noise. Essentially, the nodes are adapted, provided that the standard errors due to the adaptation result in a statistically significant improvement, typically determined through an F-test. We assume S polynomial sections in the whole sequence u. The s^{th} cubic polynomial is given by

$$u_{s-1}(\sigma) = a_s + b_s\sigma + c_s\sigma^2 + d_s\sigma^3, \tag{1}$$

where σ is the local time of the section. For a systematic approach, we write the equations in matrix form as

$$V_s\theta_s = U_s, \tag{2}$$

where $\theta_s = [a_s, b_s, c_s, d_s]^T$ is the unknown vector, which has 4 unknown coefficients, V is a Vandermonde matrix computed from instants in the section. We assume K time samples in this section.

$$V_s = \begin{bmatrix} 1 & \sigma_1 & \sigma_1^2 & \sigma_1^3 \\ 1 & \sigma_2 & \sigma_2^2 & \sigma_2^3 \\ \vdots & \vdots & \cdots & \vdots \\ 1 & \sigma_K & \sigma_K^2 & \sigma_K^3 \end{bmatrix}. \tag{3}$$

For the whole sequence, we have

$$C\theta = U, \tag{4}$$

where C collects the Vandermonde matrices of all sections:

$$C = \text{diag}\,(V_1, \cdots, V_S), \tag{5}$$

and the unknown θ represents all coefficients of all sections (in MATLAB™ notation):

$$\theta = \left[\theta_1; \quad \theta_2; \quad \cdots; \quad \theta_S\right]. \tag{6}$$

Note that the least-squares solution minimizes

$$\min_{\theta} \; ||C\theta - U||^2 = \sum_j \left(V^j\theta - U^j\right)^2.$$

Now assume the data are measured at times $t_1, \cdots, t_n$ and that there are S sections defined by the nodes $\tau_1, \cdots, \tau_{S+1}$. We assume that the sample times are split according to

$$\begin{aligned} &\tau_1 \le t_1 < t_2 < \cdots < t_{K_1} < \tau_2, \\ &\quad\vdots \\ &\tau_S \le t_{K_{S-1}+1} < \cdots < t_{K_S} \le \tau_{S+1}. \end{aligned}$$

In this case, every sample $\left(t_j, u\left(t_j\right)\right)$ is counted/used once. The data is gathered in $U = \left[U_1; U_2; \cdots; U_S\right]$ represents the measurement temporal sequence u

$$U_1 = \begin{bmatrix} u_1 \\ \cdots \\ u_{K_1} \end{bmatrix}, \ldots, \quad U_S = \begin{bmatrix} u_{K_{S-1}+1} \\ \vdots \\ u_{K_S} \end{bmatrix}. \tag{7}$$

In order to make the approximation smooth and connected, the polynomials and their derivatives must be continuous as the connecting points. The continuity condition and the smoothness conditions between sections s and $s+1$ are

$$\begin{aligned} \text{continuity:} \quad & u_s(w_s) = u_{s+1}(0), \\ \text{smoothness:} \quad & u_s'(w_s) = u_{s+1}'(0), \end{aligned}$$

which can be written in terms of the unknown coefficients as

$$\begin{aligned} a_s + b_s w_s + c_s w_s^2 + d_s w_s^3 &= a_{s+1}, \\ 0 + b_s w_s + 2c_s w_s + 3d_s w_s^2 &= b_{s+1}. \end{aligned} \tag{8}$$

A matrix used to implement the continuity constraints between sections s and $s+1$ is introduced and denoted G_s

$$G_s = \begin{bmatrix} 1 & w_s & w_s^2 & w_s^3 & -1 & 0 & 0 & 0 \\ 0 & 1 & 2w_s & 3w_s^2 & 0 & -1 & 0 & 0 \end{bmatrix}, \tag{9}$$

where w_s is the length of sth section.

For the whole sequence, the LHS of the constraints is represented

$$A_{eq} = \begin{bmatrix} \boxed{G_1} & 0 & \ldots & 0 \\ 0 & \boxed{G_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \boxed{G_{S-1}} \end{bmatrix}. \tag{10}$$

Note that G_s and G_{s+1} overlap over 4 columns. The RHS of the constraints is represented as a column of $2(S-1)$ zeros.

Once the LHS is formed, a cubic-spline fit is applied to the data given an initial grid of points. Subsequently the grid is adapted so as to make the residual errors as uniform and homoscedastic as possible. This involves three operations:

a) Inserting nodes: This operation is based on the null hypothesis that the variances of two residuals (before and after inserting a node) are equal. If the null hypothesis is rejected, (namely, that the F-statistic is significantly larger than 1), the current section can be divided into smaller sections by inserting a node. Note that in this way the heteroscedasticity of the residual error is reduced, and the estimated trend leaves a hopefully homoscedastic residual.

b) Shifting nodes: The shifting is simply determined by whether it improves the overall standard error. The shifting is first tested in the positive direction, and if it fails to improve the error, the current node is shifted in the negative direction.

c) Removing nodes: This is similar to the operation of inserting nodes, and is based on the null hypothesis that the variances of two residuals (before and after inserting a node) are equal. If the null hypothesis is rejected, (namely, that the F-statistic is significantly larger than 1), the two sections in questions are merged by removing the node between them. Heteroscedasticity of the residual error is therefore hopefully maintained.

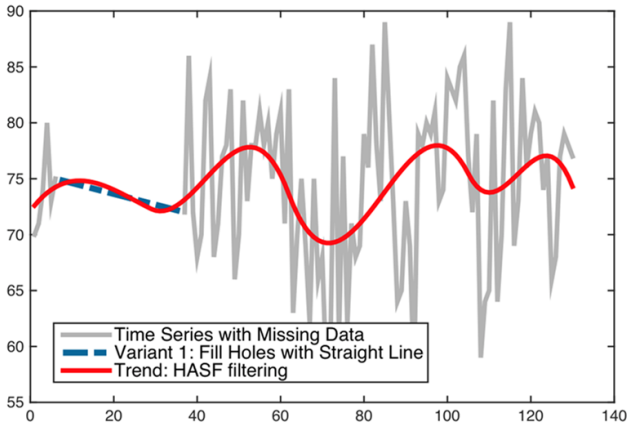

FIGURE 3. Variant 1: A missing data gap is filled by a straight line defined by the data at its endpoints.

B. MODIFICATIONS OF HASF

1) VARIANT 1: FILLING GAPS WITH STRAIGHT LINES

In order to initialize a complete time series, we use straight lines to connect the gaps of fragmented segments. In Figure 3, the gray color signal is the time series with missing data; the blue color line is the dummy data that fills in the gap; the red color line is the trend of the filled time series using HASF filtering. Note that the accuracy of straight line filling is dependent on the edge points of the gap of missing data.

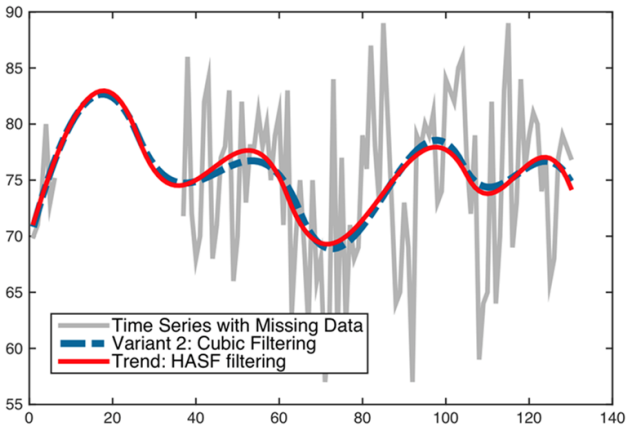

FIGURE 4. Variant 2: Missing data are replaced with cubic splines, and data-dense regions are approximated with cubic splines.

Moreover, the fitted cubic splines (over existing data) and the straight gap-filling line are discontinuously connected.

2) VARIANT 2: FILLING GAPS WITH CUBIC FILTERING

In this variant, we use cubic splines to fill the gaps with interpolated data (Figure 4). If a data point is missing, we remove the row of U and the corresponding row of C. Here continuity and smoothness are guaranteed between the polynomials fitting the existing data and the gaps. Unfortunately, the cubic polynomials over the gaps are expected to suffer from wild oscillations.

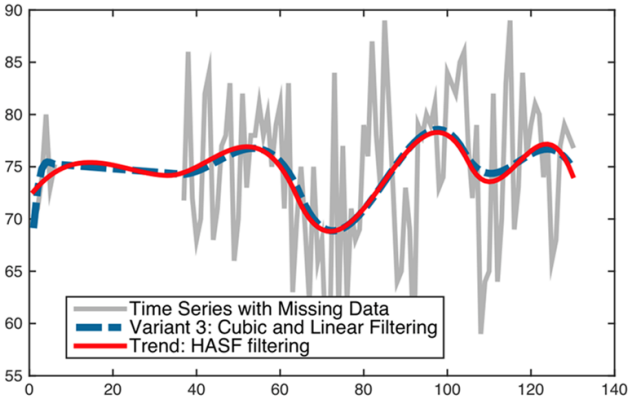

FIGURE 5. Variant 3: Missing data are replaced with linear splines while data-dense regions are approximated with cubic splines.

3) VARIANT 3: FILLING GAPS WITH COMBINED CUBIC AND LINEAR FILTERING

In this variant, we use linear splines to fill the data gaps and cubic splines to fit the existing data, and we enforce continuity and smoothness at every node. Note that gaps that are too short, for instance, single points surrounded by existing samples, are ignored in the sense that the section containing them is still represented with a cubic spline fitted using the existing data. The result is illustrated in Figure 5. The algorithm is described below.

We assume S sections in the whole sequence u. If the s^{th} section represents a gap of missing data, then a linear spline is used:

$$u_{s-1}(\sigma) = a_s + b_s\sigma,$$

where σ is the "local" time of the section considered. The measurement equations are given by

$$V\theta_s = u, \tag{11}$$

where $\theta_s = [a_s, b_s]^T$ is the unknown vector, which has 2 unknown coefficients. The corresponding Vandermonde matrix V of a missing data segment is

$$V_s = \begin{bmatrix} 1 & \sigma_1 \\ 1 & \sigma_2 \\ \cdots & \cdots \\ 1 & \sigma_K \end{bmatrix}. \tag{12}$$

In order to make the approximation smooth and connected, the corresponding functions and the derivatives should be continuous at the points connecting the sections. The continuity condition $u_s(w_s) = u_{s+1}(0)$ between sections s and $s+1$ is modified to

$$a_s + b_s w_s = a_{s+1}, \tag{13}$$

where w_s is the length of s^{th} section. The smoothness condition $u'_s(w_s) = u'_{s+1}(0)$ becomes

$$0 + b_s w_s = b_{s+1}, \tag{14}$$

A matrix used to implement the continuity constraints between sections s and $s+1$ is introduced and denoted G_s. There are 4 constraints included below:

- If section s uses linear filtering and section $s+1$ uses cubic filtering, then
$$G_s = \begin{bmatrix} 1 & w_s & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \end{bmatrix}. \tag{15}$$
G_s and G_{s+1} overlap over 2 columns in G.
- If section s uses cubic filtering and section $s+1$ uses linear filtering, then
$$G_s = \begin{bmatrix} 1 & w_s & w_s^2 & w_s^3 & -1 & 0 \\ 0 & 1 & 2w_s & 3w_s^2 & 0 & -1 \end{bmatrix}. \tag{16}$$
G_s and G_{s+1} overlap over 4 columns in G.
- If section s uses cubic filtering and section $s+1$ uses cubic filtering, then
$$G_s = \begin{bmatrix} 1 & w_s & w_s^2 & w_s^3 & -1 & 0 & 0 & 0 \\ 0 & 1 & 2w_s & 3w_s^2 & 0 & -1 & 0 & 0 \end{bmatrix}. \tag{17}$$
G_s and G_{s+1} overlap over 4 columns in G.
- If section s uses linear filtering and section $s+1$ uses linear filtering, then
$$G_s = \begin{bmatrix} 1 & w_s & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}. \tag{18}$$
G_s and G_{s+1} overlap over 2 columns in G.

All these constraints can be obtained by appropriate deletions of the columns of the original G_s.

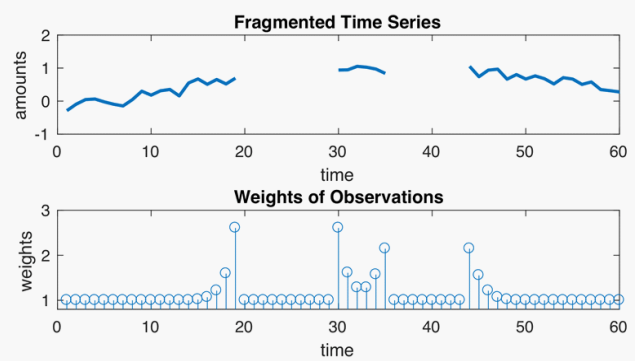

FIGURE 6. An example of adding weights to increase the importance of data neighboring missing data.

4) IMPORTANCE-WEIGHTING OF THE MEASUREMENT EQUATIONS

In the case of missing data, it is better to use the weighted least squares criterion. A good strategy is to assign a higher weight to the observation points that are isolated. If a point is surrounded by missing observations, that point should be counted as more important; in other words, it should possess a higher weight, as illustrated in Figure 6. Note that the weights of the missing points are irrelevant in the least squares solution.

This is achieved by scaling the j^{th} residual, which is numerically equivalent to scaling the j^{th} rows of C and U by ρ_j:

$$\begin{aligned} \min_{\theta} \ ||C\theta - U||_\rho^2 &= \sum_j \rho_j^2 \left(C^j\theta - U^j\right)^2 \\ &= \sum_j \left(\rho_j C^j\theta - \rho_j U^j\right)^2 \end{aligned}$$

where ρ_j is a weight given to the j^{th} data sample. The algorithm of computing weights is described in Algorithm 1 below.

IV. EXPERIMENT & EVALUATION

Once the gaps of missing segments are filled, we then can apply the HASF method [10] to the filled time series to estimate the trend. The HASF method can be recursively applied as needed on the dataset to improve performance. Our experience with the examples we considered indicates, however, that one iteration is often sufficient to obtain good approximations.

To evaluate the performance of the three variants, we first applied the HASF method to the original time series without missing data to get the theoretical trend trend_c. We then computed the trend trend_m of time series with missing data. Equation 19 is used for evaluating the Signal-to-Error Ratio (SER). A higher SER indicates better performance:

$$\text{SER} = 20\log_{10}\left(\frac{\text{rms}(\text{trend}_c)}{\text{rms}(\text{trend}_c - \text{trend}_m)}\right)\text{dB} \tag{19}$$

where rms denotes the root mean square of a time series.

Algorithm 1 Algorithm: Compute Weights

u: *input a fragmented signal of length n*
w: *hamming window length*
`rho`: *importance weights*

1) *compute indices of missing observations m*
2) *compute hamming vector h*
3) *initialize weights:* `rho= ones(1, n)`
4) **loop** *over m*
   a) `lw` = *left side of current window centered at i*
   b) `rw` = *right side of current window centered at i*
   c) *find instants of missing observations* `mw` *and of existing observations* `ew` *in the current window*
   d) *compute hamming vector of current window*: `ih = (lw:rw) − i + w + 1; hw = h(ih);`
   e) `mip = sum(hw(ew));`
   f) **if** `mip > 0`
      i) `weights = hw/mip;`
      ii) `weights(mw) = weights(mw)*0;`
      iii) `rho(lw:rw)=rho(lw:rw)+weights`
5) **return** `rho`

A. HEART RATE DATASETS

The evaluation datasets are generated from Mackowiak's heart rate dataset [23], which has 130 observations in the time series. We randomly removed some observations according to two parameters, n and $w_{\max}$, where n is the number of gaps and $w_{\max}$ is the maximum length of a gap. We generated 10 datasets for each combination of n and $w_{\max}$. In total, 120 datasets were generated.

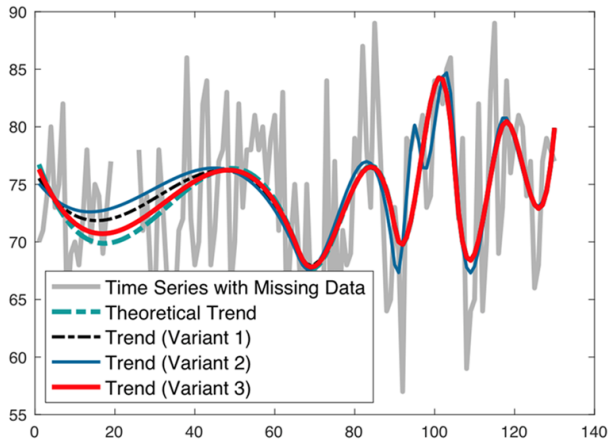

FIGURE 7. Heart Rate Dataset.

Figure 7 shows one example of experiments on the heart rate datasets. The black color line, blue color line, and red color line represent Variant 1, Variant 2 and Variant 3, respectively. The green color line is the trend of the original time series without missing data. Thus, any line that is closer to the green color line has better performance.

TABLE 2. Heart rate Datasets.

Gaps			Signal-to-Error Ratio (dB)		
$w_{\max}$	Average $w_{\max}$	n	Variant 1	Variant 2	Variant 3
2	1.6	2	32.63	25.67	34.07
5	4.3	2	28.75	24.38	30.29
10	8.1	2	27.98	16.98	28.08
15	10.3	2	27.64	14.12	28.67
20	13.3	2	27.72	12.64	29.24
25	16.9	2	26.62	7.4	27.47
2	2	5	29.52	24.32	29.47
5	4.8	5	29.06	20.78	28.71
10	8.7	5	26.28	12.51	26.75
15	13.9	5	25.41	3.49	26.03
20	16.5	5	25.27	-4.01	25.35
25	20.8	5	23.54	-8.39	24.09

We evaluated the three variants on 120 heart rate datasets. From the results of Table 2, Variant 1 (filling gaps with a straight lines) and Variant 3 (filling gaps with cubic and linear filtering) are better than Variant 2 (filling gaps with cubic filtering). Variant 3 is the best in general. In all three variants, the smaller gap, the better the SER.

TABLE 3. Blood pressure Datasets.

Gaps			Signal-to-Error Ratio (dB)		
$w_{\max}$	Average $w_{\max}$	n	Variant 1	Variant 2	Variant 3
2	1.5	2	26.33	24.26	28.76
5	4.1	2	23.63	21.62	29.99
10	7.9	2	19.37	17.15	22.22
15	11	2	17.01	14.44	20.13
20	13.9	2	15.51	13.38	18.66
25	14.5	2	14.91	13.36	17.80
2	1.9	5	23.30	22.63	28.19
5	4.9	5	18.64	18.62	20.59
10	9.2	5	16.69	9.49	19.36
15	12.3	5	14.87	7.01	17.15
20	18.5	5	11.96	8.39	12.97
25	20.1	5	11.79	3.64	12.68

B. BLOOD PRESSURE DATASETS

We also evaluated our method on a blood pressure dataset [38]. This dataset consists of time series observations of blood pressure measurements and readings from sensors. The dataset we used has 231 observations. We randomly removed some observations according to n and $w_{\max}$ as in the procedure for the heart rate datasets. Each combination (n and $w_{\max}$) has 10 datasets, and 120 datasets were generated in total. Table 3 shows the blood pressure dataset results, which are similar to those of the heart rate datasets (Table 2). As with the heart rate datasets, Variant 3 is the best, followed by Variant 1 and Variant 2, and all variants produce a better SER as the average gap length $w_{\max}$ is reduced.

C. NOISY SINE DATASETS

To study the effect of the severity of missing data, we generated a sequence of sinusoidal signals corrupted with noise,

$$u = \sin \frac{2\pi t}{P} + \text{noise} \tag{20}$$

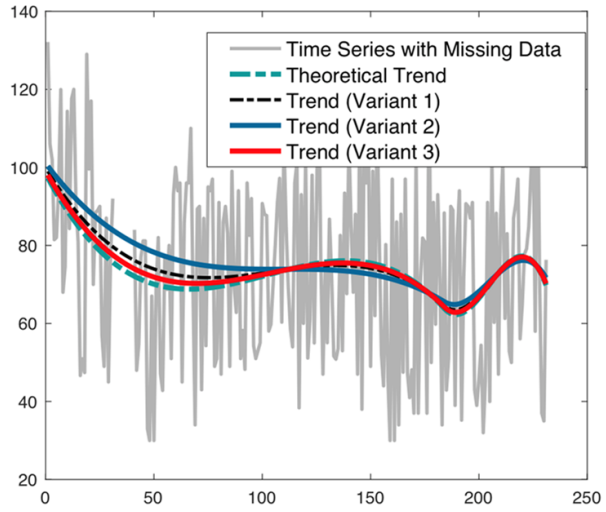

FIGURE 8. Blood pressure Dataset.

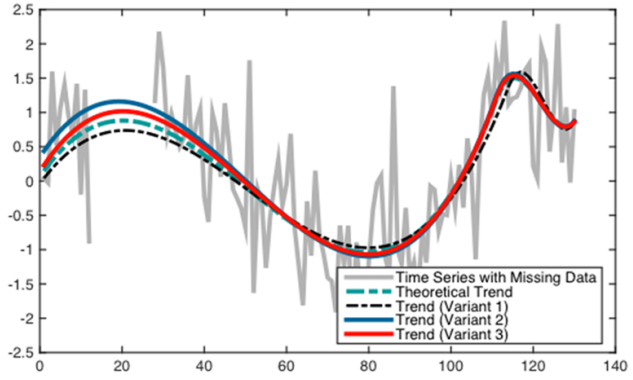

FIGURE 9. Data based on sin($2\pi t/100$)+ noise.

where P is the period of the signals, and we varied the amount of missing data. The ratio of the total length of missing data gaps normalized by the period of the underlying trend should be a good measure of the severity of gaps. In total, 120 datasets were generated and corrupted by noise (SNR=0.97).

The results of our trend detection algorithms (Table 4) applied to these data are quite consistent and similar to the heart rate datasets (Table 2) and blood pressure datasets (Table 3). Variant 3 has the best SER performance over all gap sizes $W = w_{\max}$, and the performance improves as the gap size gets smaller.

Moreover, we generated more sine datasets with different periods (50, 100, 150 and 200) and $W = w_{\max}$ (1, 3, 5, 7 and 9). Twenty combinations of period and average gap width were used in total, and each combination had 90 random datasets. We show the variation of the SER against the ratio W/P in Figure 10. An increase of W/P is likely to decrease the performance (SER). This means that a smaller gap (average $w_{\max}$) results in a better performance. Figure 11 shows that the performance is improved through importance weighting of the measurements.

TABLE 4. Sine Datasets : sin(t*2*pi/100)+noise.

Gaps			Signal-to-Error Ratio (dB)		
$w_{\max}$	Average $w_{\max}$	n	Variant 1	Variant 2	Variant 3
2	1.5	2	26.33	24.26	28.76
5	4.1	2	23.63	21.62	29.99
10	7.9	2	19.37	17.15	22.22
15	11	2	17.01	14.44	20.13
20	13.9	2	15.51	13.38	18.66
25	14.5	2	14.91	13.36	17.80
2	1.9	5	23.30	22.63	28.19
5	4.9	5	18.64	18.62	20.59
10	9.2	5	16.69	9.49	19.36
15	12.3	5	14.87	7.01	17.15
20	18.5	5	11.96	8.39	12.97
25	20.1	5	11.79	3.64	12.68

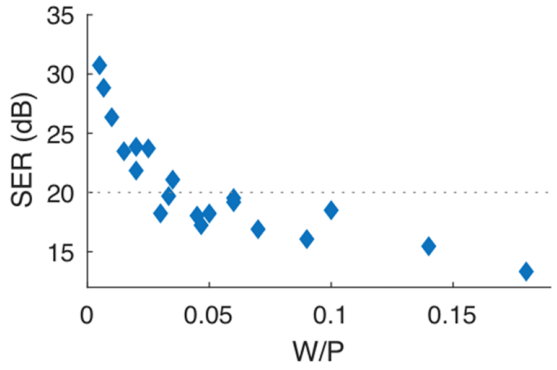

FIGURE 10. Variant 3 Performance (without importance weighting).

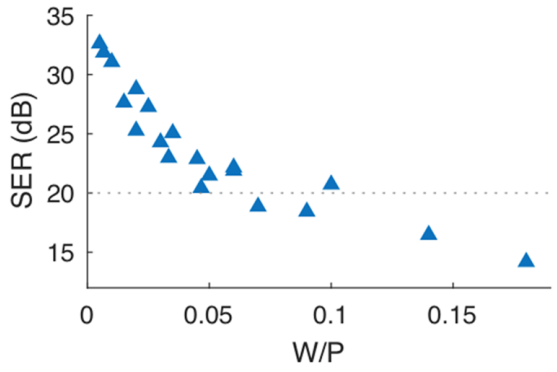

FIGURE 11. Variant 3 Performance (with importance weighting).

In summary, the performance of Variant 3 is better than that of Variant 1 and Variant 2. Overall, Variant 3 contains a number of useful characteristics, such as: (a) It ignores short gaps and models with cubic splines the data sections dominated by existing data; (b) It models data sections dominated by large gaps with linear splines; (c) It enforces continuity and smoothness conditions between all adjacent sections regardless of whether they are linear or cubic; (d) It adapts the nodes of the trend so as to improve stationarity; (e) It is ultimately augmented with importance weighting. The results are very encouraging with the signal-to-error ratio exceeding 32 dB

when the gaps are small compared with the time scale of the signal corrupted with noise. Importance weighting seems to improve the result by an average of 3.19 dB.

V. CONCLUSIONS

In this paper, we presented three variants of modifying the Hypothesis Testing Based Adaptive Spline Filtering (HASF) method to discover the trends from fragmented time series. The modified methods were shown to provide a robust way to deal with missing data due to the ability of the modified HASF to (a) ensure continuity and smoothness of the underlying trend as desired, (b) deal with the nonstationarity and heteroscedasticity of the data, and (c) reflect the change of importance of data samples that remain after removing neighboring data. The results were consistent across real and simulated datasets and show that Variant 3 is the best. Variant 3 uses cubic splines to fit the existing data and linear splines for large gaps. Importance weighting typically improves the performance with the SER exceeding 32 dB.

REFERENCES

[1] Care Continuum Alliance (CCA). *Definition of Disease Management*, Accessed on May 18, 2017. [Online]. Available: http://www.carecontinuum.org/dm_definition.asp

[2] O. Anava, E. Hazan, and A. Zeevi, "Online time series prediction with missing data," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 2191–2199.

[3] *Beddit Sleep Tracker*, accessed on Sep. 21, 2016. [Online]. Available: https://itunes.apple.com/us/app/beddit-sleep-tracker/id794968897?mt=8

[4] P. Bondon and N. Bahamonde, "Least squares estimation of ARCH models with missing observations," *J. Time Ser. Anal.*, vol. 33, no. 6, pp. 880–891 2012.

[5] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market,"*J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2010.

[6] R. Bos, S. D. Waele, and P. M. T. Broersen, "Autoregressive spectral estimation by application of the Burg algorithm to irregularly sampled data," *IEEE Trans. Instrum. Meas.*, vol. 51, no. 6, pp. 1289–1294, Dec. 2002.

[7] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1994.

[8] P. M. T. Broersen, S. de Waele, and R. Bos, "Autoregressive spectral analysis when observations are missing," *Automatica*, vol. 40, no. 9, pp. 1495–1504, Sep. 2004.

[9] O. L. V. Costa and S. Guerra, "Stationary filter for linear minimum mean square error estimator of discrete-time Markovian jump systems," *IEEE Trans. Autom. Control* vol. 47, no. 8, pp. 1351–1356, Aug. 2002.

[10] X. Dai and M. Bikdash, "Discovering disease trends using nonstationary time series of social media: Hypothesis testing based adaptive spline filtering method," *IEEE Access*, to be published.

[11] X. Dai, "Public health surveillance using social media: Short text classification and trend analysis of nonstationary time series," Ph.D. dissertation, ProQuest, Ann Arbor, MI, USA, 2017, Art. no. 10706.

[12] X. Dai, M. Bikdash, and B. Meyer, "From social media to public health surveillance: Word embedding based clustering method for twitter classification," in *Proc. SoutheastCon*, Concord, NC, USA, Mar. 2017, pp. 1–7.

[13] X. Dai and M. Bikdash, "Hybrid classification for tweets related to infection with influenza," in *Proc. IEEE SoutheastCon*, Apr. 2015, pp. 1–5.

[14] E. Damsleth, "Interpolating missing values in a time series," *Scandinavian J. Statist.*, vol. 7, no. 1, pp. 33–39, 1980.

[15] G. Eysenbach, "What is E-health?" *J. Med. Internet Res.*, vol. 3, no. 2, p. e20, 2001.

[16] *Fitbit Activity Trackers*, accessed on Jan. 5, 2017. [Online]. Available: https://www.fitbit.com

[17] *Fitness Dashboard for Apple Health App*, accessed on Jan. 5, 2017. [Online]. Available: http://flaskapp.com/fitport/

[18] M. Hadidi and S. Schwartz, "Linear recursive state estimators under uncertain observations," *IEEE Trans. Autom. Control*, vol. 24, no. 6, pp. 944–948, Dec. 1979.

[19] *HealtheConnect of Morristown Medical Center*, accessed on Jan. 5, 2017. [Online]. Available: https://itunes.apple.com/us/app/be-well-morristown-medical-center/id798745007?mt=8

[20] *Health Mate, Steps Tracker and Life Coach*, accessed on Jan. 5, 2017. [Online]. Available: https://itunes.apple.com/us/app/health-mate-steps-tracker/id542701020?mt=8

[21] *Jawbone Fitness Trackers*, accessed on Jan. 5, 2017. [Online]. Available: https://jawbone.coml

[22] X. Jiang, G. Wallstrom, G. F. Cooper, and M. M. Wagner, "Bayesian prediction of an epidemic curve," *J. Biomed. Inf.*, vol. 42, no. 1, pp. 90–99, 2009.

[23] J. Karl, *An Introduction to Digital Signal Processing*. San Francisco, CA, USA: Academic, 1989.

[24] L. K. Koivusilta, T. P. Lintonen, and A. Rimpelä, "Orientations in adolescent use of information and communication technology: A digital divide by sociodemographic background, educational career, and health," *Scandinavian J. Public Health*, vol. 35, no. 1, pp. 95–103, 2007.

[25] S. Krishna, S. Boren, and E. Balas, "Healthcare via cell phones: A systematic review," *Telemed. e-Health*, vol. 15, no. 3, pp. 231–240, 2009.

[26] W. Lee, S. J. Stolfo, and K. W. Mok, "Adaptive intrusion detection: A data mining approach," *Artif. Intell. Rev.*, vol. 14, no. 6, pp. 533–567, 2000.

[27] X. Liu and G. Andrea, "Kalman filtering with partial observation losses," in *Proc. 43rd IEEE Conf. Decision Control (CDC)*, vol. 4. Dec. 2004, pp. 4180–4186.

[28] R. H. Lewellen and S. H. Vessey, "Analysis of fragmented time series data using box-jenkins models," *Commun. Statist. Simul. Comput.*, vol. 28, no. 3, pp. 667–685, 1999.

[29] *Lose It!, Weight Loss Program and Calorie Counter*, accessed on Jan. 5, 2017. [Online]. Available: https://itunes.apple.com/us/app/lose-it!-weight-loss-program/id297368629?mt=8

[30] R. A. Mahir and A. M. H. Al-Khazaleh. (Nov. 2008). "Estimation of missing data by using the filtering process in a time series modeling," [Online]. Available: https://arxiv.org/abs/0811.0659

[31] S. Mirsaidi, G. A. Fleury, and J. Oksman, "LMS-like AR modeling in the case of missing observations," *IEEE Trans. Signal Process.*, vol. 45, no. 6, pp. 1574–1583, Jun. 1997.

[32] J. Murphy, *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications*. New York, NY, USA: New York Institute of Finance, 1999.

[33] N. Nahi, "Optimal recursive estimation with uncertain observation," *IEEE Trans. Inf. Theory*, vol. 15, no. 4, pp. 457–462, Jul. 1969.

[34] D. Nichols, "Estimation of missing data by least squares," Stanford Exploration Project, Tech. Rep. 65, Jan. 1998, pp. 262–292. [Online]. Available: http://sep.stanford.edu/data/media/public/docs/sep65/sep65/dave.ps.gz

[35] J. Nilsson, B. Bernhardsson, and B. Wittenmark, "Stochastic analysis and control of real-time systems with random time delays," *Automatica*, vol. 34, no. 1, pp. 57–64, 1998.

[36] P. A. Owili, L. Orawo, and D. Nassiuma, "Estimation of missing values for pure bilinear time series models with student-t innovations," *Int. J. Statist. Appl.*, vol. 5, no. 6, pp. 293–301, 2015.

[37] *Pacifica, Anxiety, Stress, & Depression Relief*, Accessed on Jan. 5, 2017. [Online]. Available: https://itunes.apple.com/us/app/pacifica-anxiety-stress-depression/id922968861?mt=8

[38] *Pain Prediction Data*, Accessed on Jan. 5, 2017. [Online]. Available: https://idash-data.ucsd.edu/community/49

[39] J. Parker, Y. Wei, A. Yates, O. Frieder, and N. Goharian, "A framework for detecting public health trends with twitter," in *Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Mining*, Aug. 2013, pp. 556–563.

[40] G. Phillips, L. Felix, L. Galli, V. Patel, and P. Edwards, "The effectiveness of M-health technologies for improving health and health services: A systematic review protocol," *BMC Res. Notes*, vol. 3, no. 1, p. 250. 2010.

[41] P. M. T. Broersen and R. Bos, "Time-series analysis if data are randomly missing," *IEEE Trans. Instrum. Meas.*, vol. 55, no. 1, pp. 79–84, Feb. 2006.

[42] *Pillow: The Sleep Cycle Alarm Clock for Sleep Tracking*, accessed on Sep. 21, 2016. [Online]. Available: https://itunes.apple.com/us/app/pillow-sleep-cycle-alarm-clock/id878691772?mt=8

[43] *Record by Under Armour*, accessed on Sep. 15, 2016. [Online]. Available: https://record.underarmour.com/

[44] A. V. Savkin, I. R. Petersen, and S. O. R. Moheimani, "Model validation and state estimation for uncertain continuous-time systems with missing discrete-continuous data," *Comput. Electr. Eng.*, vol. 25, no. 1, pp. 29–43, 1999.
[45] J. D. Scargle, " Studies in astronomical time series analysis. II—Statistical aspects of spectral analysis of unevenly spaced data," *Astrophys. J.*, vol. 263, no. 2, pp. 835–853, Dec. 1982.
[46] M. S. B. Sehgal, I. Gondal, and L. S. Dooley, "Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data," *Bioinformatics*, vol. 21, no. 10, pp. 2417–2423, 2005.
[47] M. G. Sefidmazgi, M. Sayemuzzaman, A. Homaifar, M. Jha, and S. Liess, "Trend analysis using non-stationary time series clustering based on the finite element method," *Nonlinear Process. Geophys.*, vol. 21, no. 3, pp. 605–615, 2014.
[48] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *J. Time Ser. Anal.*, vol. 3, no. 4, no. 1982, pp. 253–264.
[49] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. I. Jordan, and S. S. Sastry, "Kalman filtering with intermittent observations," *IEEE Trans. Autom. Control*, vol. 49, no. 9, pp. 1453–1464, Sep. 2004.
[50] *Sleep Cycle Alarm Clock*, accessed on Sep. 21, 2016. [Online]. Available: https://itunes.apple.com/app/apple-store/ id320606217?mt=8
[51] *SleepBot, Smart Cycle Alarm With Motion and Sound Tracker*, accessed on Sep. 21, 2016. [Online]. Available: https://itunes.apple.com/us/app/sleepbot-smart-cycle-alarm/id578829107?mt=8
[52] S. C. Smith and P. Seiler, "Estimation with lossy measurements: Jump estimators for jump systems," *IEEE Trans. Autom. Control*, vol. 48, no. 12, pp. 2163–2171, Dec. 2003.
[53] S. Soubeyrand, C. Morris, and E. K. Bigg, "Analysis of fragmented time directionality in time series to elucidate feedbacks in climate data," *Environ. Model. Softw.*, vol. 61, pp. 78–86, Nov. 2014.
[54] *Statista, Statistics and Facts About mHealth*, accessed on Dec. 14, 2016. [Online]. Available: https://www.statista.com /topics/2263/mhealth/
[55] *Pedometer, Step Counter Activity Tracker*, accessed on Sep. 15, 2016. [Online]. Available: https://itunes.apple.com/us/app/stepsapp-pedometer-step-counter/id1037595083?mt=8
[56] A. R. Tomé and P. M. A. Miranda, "Piecewise linear fitting and trend changing points of climate parameters," *Geophys. Res. Lett.*, vol. 32, no. 2, pp. 1–4, 2004.
[57] D. de Waele and P. M. S. Broersen, "Error measures for resampled irregular data," *IEEE Trans. Instrum. Meas.*, vol. 49, no. 2, pp. 216–222, Apr. 2000.
[58] Z. Wang, D. W. C. Ho, and X. Liu, "Variance-constrained filtering for uncertain stochastic systems with missing measurements," *IEEE Trans. Autom. Control*, vol. 48, no. 7, pp. 1254–1258, Jul. 2003.
[59] Z. Wang, F. Yang, D. W. C. Ho, and X. Liu, "Robust finite-horizon filtering for stochastic systems with missing measurements," *IEEE Signal Process. Lett.*, vol. 12, no. 6, pp. 437–440, Jun. 2005.

XIANGFENG DAI received a M.S. degree from the Department of Computer Science, Durham University, U.K., and a Ph.D. degree in computation science and engineering from North Carolina A&T State University. He is currently a Data Scientist at the Institute for Medical Research, Durham VA Medical Center.

MARWAN BIKDASH received the M.S. and Ph.D. degrees in electrical engineering from Virginia Tech in 1990 and 1993, respectively. He is currently a Professor and the Chair of the Department of Computational Science and Engineering, North Carolina A&T State University. He teaches and conducts research in signals and systems, computational intelligence, and modeling and simulations of systems with applications in health, energy, and engineering. He has authored over 130 journal and conference papers. He has supported, advised, and graduated over 50 M.S. and Ph.D. students. His projects have been funded by the Jet Propulsion Laboratory, Defense Threat Reduction Agency, Army Research Lab, NASA, National Science Foundation, the Office of Naval Research, Boeing Inc., Hewlett Packard, National Renewable Energy Laboratories, the Army Construction Engineering Research Laboratory, and others.