Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–20
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280216633248
smm.sagepub.com
SAGE

Missing value imputation for physical activity data measured by accelerometer

Jung Ae Lee[1] and Jeff Gill[2]

Abstract

An accelerometer, a wearable motion sensor on the hip or wrist, is becoming a popular tool in clinical and epidemiological studies for measuring the physical activity. Such data provide a series of activity counts at every minute or even more often and displays a person's activity pattern throughout a day. Unfortunately, the collected data can include irregular missing intervals because of noncompliance of participants and therefore make the statistical analysis more challenging. The purpose of this study is to develop a novel imputation method to handle the multivariate count data, motivated by the accelerometer data structure. We specify the predictive distribution of the missing data with a mixture of zero-inflated Poisson and Log-normal distribution, which is shown to be effective to deal with the minute-by-minute autocorrelation as well as under- and over-dispersion of count data. The imputation is performed at the minute level and follows the principles of multiple imputation using a fully conditional specification with the chained algorithm. To facilitate the practical use of this method, we provide an R package *accelmissing*. Our method is demonstrated using 2003–2004 National Health and Nutrition Examination Survey data.

Keywords

Accelerometer, physical activity, missing count data, multiple imputation, zero-inflated model, Poisson log-normal

1 Introduction

Accelerometers, also called activity monitors, are useful tools for measuring physical activity of subjects in clinical and epidemiological settings. While there has been growing interest in investigating the relationship between physical activity and health outcome,[1] how to accurately measure the physical activity has been controversial. Self-report measurements, collected from diaries, questionnaires and interviews, have suffered from a variety of mis-measurements,

[1]Division of Public Health Sciences, Department of Surgery, Washington University School of Medicine in Saint Louis, Saint Louis, MO, USA
[2]Division of Public Health Sciences, Department of Surgery, School of Medicine and Department of Political Science, College of Arts and Sciences, Washington University in Saint Louis, Saint Louis, MO, USA

Corresponding author:
Jung Ae Lee, Division of Public Health Sciences, Department of Surgery, Washington University School of Medicine in Saint Louis, Saint Louis, MO 63110, USA.
Email: leeju@wudosis.wustl.edu

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

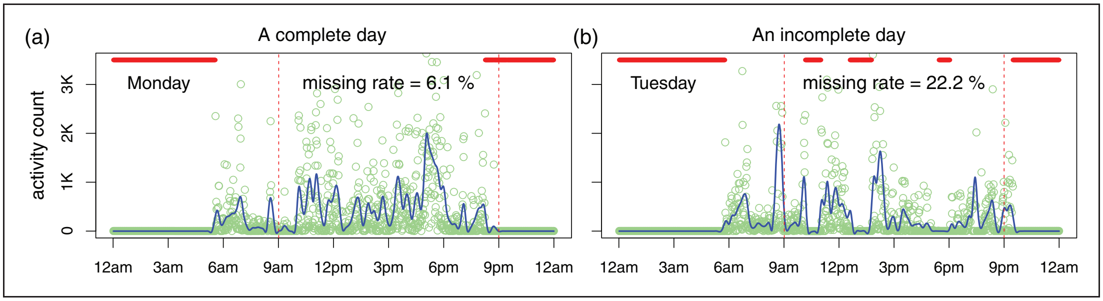

Figure 1. Illustration of physical activity pattern measured by accelerometer. The green dots are the observed accelerometer counts. The smoothed fit (blue line), produced by a B-spline, is overlayed for visual convenience. The red lines on top of each display indicate missing intervals. Panel (a) provides an example of a complete day that has a missing rate during 9 am–9 pm that is 6.1% ($\leq$ 10%) of that period. Meanwhile, panel (b) displays an example of an incomplete day that has missing rate that is 22.2% ($>$10%) during the same period.

primarily due to recall-biases or the subject's desire to meet certain social norms.[2,3] As an alternative that is free from these potential sources of errors, accelerometer use has steadily increased since 1990s as a way to objectively assess human behavior(s) in daily life. This device is usually mounted on the hip or wrist and produces a "count" value by integrating acceleration signals over every second, every minute, or a researcher-specific epoch. The resulting output depicts a series of activity counts over time that reveals a person's pattern of activity throughout the day (Figure 1).

Researchers in medicine, public health, kinesiology, and other fields find that accelerometers are important tools for measuring physical activities outside of a clinical environment. For instance, post-radical prostatectomy patients are routinely told that exercise reduces their post-operative periods of incontinence and impotence. Despite this admonition, compliance rates are lower than expected and accelerometers provide an objective measure for physicians to assess their patient's exercise rate and prospects for improvement in urinary and sexual conditions. Classically, accelerometers have been extensively used for overweight and obese patients during supervised weight loss, including those receiving bariatric surgery. There are many more examples, but the common purpose is to gain more insight into the activities of subjects during unsupervised times.

Despite the common perception that the accelerometer is an extremely accurate device, it can also produce fairly noise datasets from a statistical perspective. A second problem is missing data resulting from noncompliance of the participants; the device is sometimes removed from the hip or wrist and no motion is recorded. This occurs during sanctioned off-time such as baths, showers, swimming, but other periods of missing are deliberate and can bias subsequent analyses. For this reason, existing studies recommend to include only valid days that have sufficient wearing time.[4–7] As a result, there have been many efforts to determine the optimal criteria for the valid datasets.[8–13] However, these data reduction criteria can cause significant reduction in sample size [14] and introduce unwanted variance in subsequent estimations.[15,16]

The purpose of this study is to provide a statistical method to impute ("fill-in") missing accelerometer data. Suppose we observe multivariate count data in an $N \times T$ (typically $N < T$) accelerometer dataset, where N is the total number of days and T is the total time points in a single day. Such data are characterized by more zero values than typically predicted with standard count models ("zero-inflation") and a time-associated effect, whereby consecutive observations are highly correlated ("autoregressive covariance"). To accommodate these characteristics, we specify an imputation model based on a mixture of zero-inflated Poisson and Log-normal (ZIPLN)

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

distributions and demonstrate its efficacy with simulated and real public health data. At each indexed minute in the data, we assume a univariate missing variable (e.g., activity count at 9:01 a.m.) with complete covariates, followed by *multiple imputation with chained equations*[17] to fill-in missingness.

The innovation of this work is two-fold. *First*, the developed method is a new type of multiple imputation that is well suited to autocorrelated multivariate count data. This approach can be also viewed as an extension to the well-known fully conditional specification,[17] but incorporating a new predictive distribution adapted to the unique structure of accelerometer data. *Second*, for applied researchers working with accelerometer data, the imputation approach for irregular missing intervals is easy to implement with our R package and conveniently substitues for the dominant, but deeply flawed, case-wise deletion process. Thus, we expect this tool to provide the practical guidance to a wide range of data users.

This paper is organized as follows. In Sections 2 and 3, we display a motivating example where we discuss the missing value definition in accelerometer data. In Section 4, we propose a predictive model suitable for the activity data peculiarity. In Sections 5–7, the imputation procedure based on the ZIPLN is introduced along with some discussion.

2 Accelerometer missing data

In this section, we describe in detail the structure of missingness in accelerometer data and the associated methodological challenges.

2.1 Missing value definition

There are two ways to define missing values in accelerometer data: *missing in days* and *missing at time*. Although these two approaches are related in the sense that the missing days are determined by a certain amount of missing time per day, the two definitions have totally different interpretations in terms of the imputation process. The *missing in days* approach leads to an imputation model that provides summary statistics per day. For this approach, we first have to select the complete days that contain sufficient wearing time, e.g., more than 10.8 h a day (90% wearing during 9 a.m.−9 p.m.). This standard automatically classifies the incomplete days that have less wearing time than 10.8 h, and these incomplete days are considered missing data (Figure 1(b)). A conventional imputation method is then applied to fill in the total amount of physical activity or moderate-to-vigorous physical activity on that missing day. This process involves an inference based on the corresponding statistics of the complete days using the EM Algorithm or multiple imputation (MI).[4,5,18–22]

Table 1. Missing rate during 9 am−9 pm. 16 incomplete days ($\geq$ 10% missing rate) will be dropped, resulting in 41.71% total missing rate by days, whereas the total missing rate by minutes is only 12.77%.

(Unit: %)	Sun	Mon	Tue	Wed	Thu	Fri	Sat
Subject1	~~10~~	6	~~22~~	~~12~~	0	0	~~17~~
Subject2	0	0	0	0	~~21~~	~~41~~	~~11~~
Subject3	0	0	0	3	~~29~~	0	5
Subject4	0	~~78~~	1	0	~~11~~	~~69~~	~~13~~
Subject5	~~28~~	~~22~~	~~27~~	2	3	~~15~~	2

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

A drawback with *missing in days* imputation is that it ignores the fact that the missing time per day almost always varies across days. By simply dropping the incomplete days, one may lose some valuable information as well as overestimate the total missing rate. The example in Table 1 shows that the missing rate during the standard measurement day (9 a.m.–9 p.m.) varies across days and subjects. Suppose we determine the incomplete days by the criterion of a $\geq$10% missing rate. The total number of incomplete days for all five subjects in this example is 16 $(4+3+1+4+4)$, resulting in a total missing rate that is 41.71%, which is calculated by the number of incomplete days divided by total days. Compared with this missing day approach, the total missing rate by minute is considerably lower than that, only 12.77%, which is calculated by the amount of missing minutes (or total length of the missing intervals) divided by total minutes during 9 a.m–9 p.m. Therefore, we find that a minute level imputation method is more appealing than a day level imputation in that we can minimize the data loss as well as maximize information that we partially have per day.

2.2 Missing interval

Here, we define the missing data as the missing counts per minute, which forms an irregular time interval, rather than the incomplete day or the regularly segmented time period. One effective way to detect the missing interval is to find some extended period of minutes where exact zeros succeed more than $P = 20$, 30, or 60 min.[4,5,22,23] In the example of Figure 1 and Table 1, we defined the missing interval by sustained zeros over 20 min. Figure 2(a) shows the missing and non-missing regions with bright (yellow) and dark (red) colors, respectively, by this criterion. It is not surprising that there are many more zero counts during the obvious sleeping time. Our goal in this work is to impute the missing intervals flagged by yellow in (a).

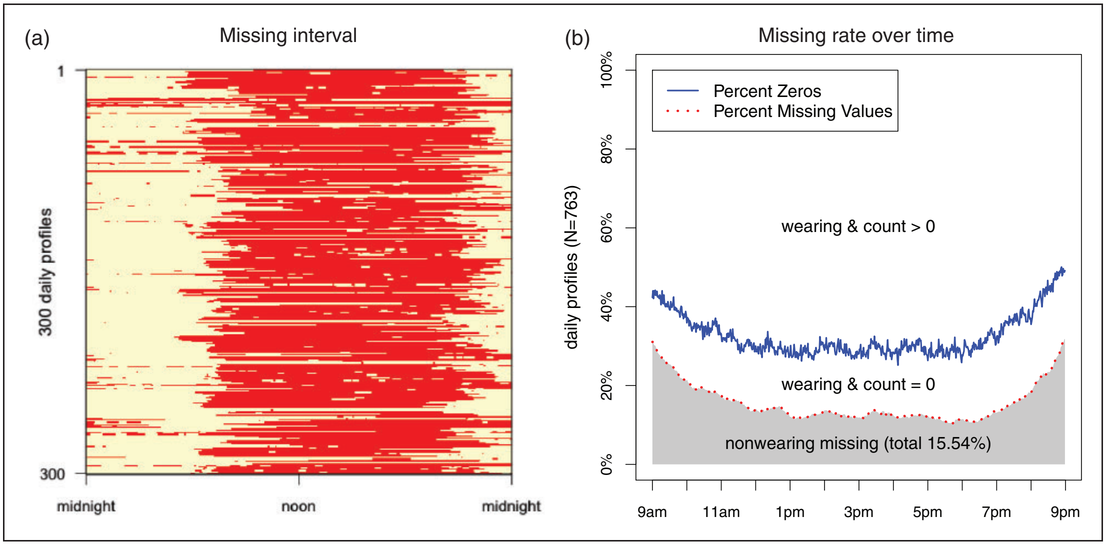

Figure 2. Illustration of missing interval definition. In panel (a), we define the missing interval by consecutive zeros more than 20 min, which is flagged by the brighter color (yellow). In panel (b), the blue solid line indicates the percent of zeros at each minute, and the red dotted line indicates the percent of non-wearing missingness. The shaded area under the dotted line integrates to the overall missing rate of 15.54%.

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

After defining the missing interval, the data set consists of three types of values: (1) positive counts, indicating some physical activities, (2) zero counts continuing less than or equal P minutes, indicating no-movement, (3) zero counts continuing more than P minutes, indicating non-wearing missing time. Figure 2(b) describes the percent of these values at each minute. For instance, at 3 p.m., the zeros are 28.3% out of a sample size of 763, and the missing values are 11.8% since only partial zeros are categorized as missing values. Notice that the total missing rate from the shaded region in Figure 2(b) is 15.54% during 9 a.m. to 9 p.m. Our goal is to impute the missing activity counts in the shaded area on the basis of information that we can obtain from the remaining 84.46% wearing time.

Accelerometer produced datasets are characterized by many zeros and autocorrelated non-negative counts in T dimensions, where T is the total time points in a day. If the accelerometer is set to record data at 1 min epoch then $T = 1440$. Zero-inflation during the wearing time occurs because there are always left-out zeros (short term zeros less than P minutes) after sorting out the missing interval by the P minutes criterion. Thus, the zero-inflation becomes more severe with a $P = 60$ min criterion than with a $P = 20$ min criterion.

The minute-level measurement is important in the study of physical activity and exercise. Researchers in this area are interested in not only the total amount of physical activity but also the activity bouts, i.e., how long the activity last. The minute level imputation makes it possible to preserve the accurate assessment of the rate and duration of such as exercise, sedentary behavior, and sleep. To date, relatively little work has been done on the imputation using minute level information. Morris et al.[24] stochastically imputes the missing METs (i.e., metabolic equivalent) during the irregular missing time interval, using a wavelet-based functional mixed Bayesian model. This method assumes that the data are missing completely at random (MCAR, the distribution of the missing data does not depend on other data missing or observed) and require a multivariate normal assumption to randomly draw the missing-wavelet coefficients. Lee[14] imputes counts per minute taking into account the available information from the invalid days and then showed that the combined method from both valid and invalid days improves the missing value imputation performance.

3 Activity data example

To illustrate the challenges of missingness in accelerometer data, we use 2003−2004 National Health and Nutrition Examination Survey (NHANES) dataset available at the website: *http://wwwn.cdc.gov/nchs/nhanes/search/nhanes03_04.aspx*. From 7176 total participants in the physical activity survey, we randomly select 218 individuals to give 1526 daily profiles (218 × 7 days). Following Catellier et al.,[4] "non-wearing time" is defined by successive zeros over 20 min. A "standard measurement day" is 9 a.m. to 9 p.m., during which over 60% of the sample wore the device. Other periods have been proposed in other literatures, but we find that these times represent sharp aggregate cutoffs in the NHANES data. Furthermore, a "complete day" is defined as a day in which a subject wears the accelerometer over 90% of the standard measurement day (at least 10.8 h). Requiring this standard reduces the number of daily profiles considered to 576. At least three complete days of a person are recommended for the reliable estimate of habitual physical activity.[12] Therefore, some complete days are dropped if the subject has only one or two completed days out of seven days, although some incomplete days are included if the subject has already three or more complete days. So, 763 daily profiles of 109 people (109 × 7 days) remain. Note that 15.54% missingness remains, as shown in Figure 2(b), after such processing. Table 2 summarizes the characteristics of the final dataset. From the table, we see that the mix of demographics is nationally representable.

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

Table 2. Summary of data (No. of participants = 109, N = 763 days).

Age(%)[a]	Sex (%)	BMI (%)	Race (%)
Youth 38.5%	Male 50.5%	$\leq$25 42.2%	White 44.0%
Adult 61.5%	Female 49.5%	>25 57.8%	Others 56.0%

[a] Youth indicates 7–19 yrs and Adult indicates 20–85 yrs.

4 Proposed method

4.1 Assumptions for MI

Our primary goal is filling-in the missing interval over $T = 1440$ min of a day with plausible count values. We do this by drawing imputations for non-wearing time from a specified distribution, a probabilistic estimate of what would have likely occurred if the subject had worn the device. Therefore, the predictive model is first built on actual wearing time, then the imputation process is carried out for non-wearing time based on information gained from the wearing time. Multiple imputation (MI) is known to be practically useful for doing this task under missing at random (MAR), (the distribution of the missing data does not depend on the other missing data) or missing at completely at random (MCAR) assumptions. In this paper, we assume that the wearing times and the non-wearing times are not fundamentally different and there are no ‘‘shocks’’ in the non-wearing periods. This assumption is more convenient than optimal and is equivalent to the MAR assumption in standard missing data analysis. It is possible for accelerometer data to be NMAR: the distribution of the missing data depends on other missing data.[25,26] Thus far, there is no general test for the missing mechanism MAR (or MCAR) versus NMAR. In fact, with any method, the NMAR mechanism is hard to specify. Therefore, our method takes advantage of the simplicity of MI model under MAR and demonstrates the reasonablility of MAR assumption (Section 6 and Appendix 1.1). Handling NMAR missing accelerometer data is a separate challenge from that addressed here.

There are two problems with directly applying the standard MI process to accelerometer missing data. First, the probability model specified by normal or other standard distribution is not suitable for minute level activity count data. Second, the high dimensionality of accelerometer data $(N < T)$ where some of dimensions are strongly correlated makes it hard to use any method that is developed for $N > T$ rectangular multivariate data. Our method provides a way to obtain a suite of complete datasets tailored to these challenges, where researchers are then able to perform the needed analysis and combine the results in the usual second part of the MI process. It is also important to note that the main value of MI stems from its wide applicability and production of unbiased coefficient estimators under MAR with standard rectangular data structures. It is usually not *optimal* in the sense that customized imputation schemes that exploit unique structures in particular datasets and sometimes impose additional distributional assumptions are likely to produce the smaller standard deviations for a column with missing data and smaller standard errors for the associated coefficient estimate. This is the philosophical basis for our approach to improving imputation schemes for accelerometer missing data.

4.2 Zero-inflated Poisson model

Zero-inflated Poisson (ZIP) regression was first introduced in Lambert[27] although the ZIP distribution, without covariates, had been discussed earlier elsewhere.[28,29] The main purpose of

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

this approach is to deal with so called "structural" zeros in modeling count data that exceed those predicted by a regular Poisson generalized linear model (GLM). The same principle is applied to handle the physical activity counts data that often encounter excess zeros resulting from frequent no-movement by subjects. The ZIP regression model assumes that zeros are observed with probability π, and the rest of observations come from a Poisson(λ) with probability $1-\pi$. Let $Y_1, \ldots, Y_N$ be a sample of size N independently drawn from

$$Y_i \sim \begin{cases} 0 & \text{with probabilit}y\ \pi_i \\ \text{Poisson}(\lambda_i) & \text{with probability } 1-\pi_i \end{cases}$$

and the probability mass function is given by

$$P(Y_i = h) = \begin{cases} \pi_i + (1-\pi_i)e^{-\boldsymbol{\lambda}_i} & \text{for } h = 0 \\ (1-\pi_i)e^{-\lambda_i}\lambda_i^h/h! & \text{for } h = 1, 2, \ldots. \end{cases} \tag{1}$$

Accordingly, the regression model with this ZIP distribution consists of two GLM components. The first part is a logistic regression, specified by $\text{logit}(\pi_i) = \boldsymbol{u}_i^T\boldsymbol{\gamma}$, where the response variable states zero or nonzero status and $\boldsymbol{\gamma}$ is a regression coefficient vector for covariates $\boldsymbol{u}_i^T$. The second part is a Poisson regression, specified by $\log(\lambda_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$, where the response variable is a non-negative count from a Poisson(λ_i) and $\boldsymbol{\beta}$ is a regression coefficient vector for covariates $\boldsymbol{x}_i^T$. This separation allows the predictors in each model to perform different roles; for example, what causes exact zeros (no-movement) is different from what causes vigorous activities. For these covariates, we can consider age, sex, race, body mass index, weekday vs. weekend, and others. The mean and variance of ZIP distribution are

$$\text{E}(Y_i) = (1 - \pi_i)\lambda_i, \qquad \text{Var}(Y_i) = \lambda_i(1 - \pi_i)(1 + \pi_i\lambda_i)$$

where we can see that when $\pi \to 0$, i.e., zero-inflation disappears, the mean and variance become the same, holding the mean-variance relationship of standard Poisson distribution.

A similar approach to handle zero-inflated count data is also introduced in Mullahy[30] and is referred to as a hurdle model. This model utilizes a zero-truncated Poisson distribution, i.e., $P(Y_i = h \mid Y_i > 0) = \lambda^h/((e^{\lambda_i} - 1)h!)$. Thus, the probability mass function in equation (1) is modified to

$$P(Y_i = h) = \begin{cases} \pi_i & \text{for } h = 0 \\ (1-\pi_i)\lambda_i^h/((e^{\lambda_i} - 1)h!) & \text{for } h = 1,\ 2, \ldots. \end{cases}$$

The hurdle model provides results that are nearly identical to the ZIP model in our empirical results, so we choose to use the more intuitive ZIP model.

4.3 Multivariate ZIP regression on activity data

Again consider an N by T data matrix, where N is the total number of daily profiles and T is the total number of time points in a day; $T = 1440$ for a 1 min epoch. We define two sets of indices $\mathcal{N} = \{1, \ldots, N\}$ and $\mathcal{T} = \{1, \ldots, T\}$ and denote $i \in \mathcal{N}$ and $t \in \mathcal{T}$. Because the wearing versus non-wearing status of N observations varies across time, as seen in Figure 2(b), it is necessary to partition $\mathcal{N}$ into two parts, $\mathcal{N} = \mathcal{W} \cup \mathcal{W}^c$, by creating a subset $\mathcal{W}$ that only includes indices of the wearing status, while $\mathcal{W}^c$ is the complement set of $\mathcal{W}$ containing the non-wearing profile indices. The elements of $\mathcal{W}$ and $\mathcal{W}^c$ change as time proceeds (we omit the subscript t in some situations for

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

simplicity). Furthermore, assume that there were no movements if the subjects did not wear the device during regular sleeping time. This narrows the focus on only the daytime hours, 9 a.m. to 9 p.m., so the domain is $\mathcal{D} = \{541, \dots, 1260\}$ where $\mathcal{D} \subset \mathcal{T}$. Therefore

$$P(Y_{it} = 0 \mid i \in \mathcal{W}_t^c \text{ and } t \in \mathcal{D}^c) = 1 \tag{2}$$

Because of assumption (2) that treats the missingness outside of the domain $\mathcal{D}$ as the extended sleep period, our method may not be useful for a sleep study that requires sensitive detection of the onset of the sleep/wake-up time. Thus, we recommend that our method be used primarily for the studies focused on daytime activity(or inactivity).

The prediction from ZIP regressio is based on the expected value of a ZIP random variable in equation (2), which is expressed as a conditional expectation in the regression context by substitution:

$$\mathrm{E}(Y_i | \boldsymbol{u}_i, \boldsymbol{x}_i) = \left\{1 - \mathrm{logit}^{-1}(\boldsymbol{u}_i^T \gamma)\right\} \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right), \qquad i \in \mathcal{W} \tag{3}$$

where $\boldsymbol{u}_i$ and $\boldsymbol{x}_i$ are vectors of covariates for logit and Poisson model, respectively, and γ are the corresponding coefficient vectors. Now, we consider the multivariate setting where the ZIP regression is applied for multiple time points $t \in \mathcal{D}$. An important consideration for this model is how to incorporate the time dependency. It makes sense in serial physical activity data that the level of motion in the current minute is correlated with those that follow. Therefore, modify the regression model (3) by adding autoregressive terms $H_{i,t-1}$ as follows

$$\mathrm{E}\left(Y_{i,t} | Y_{i,t-1,} \, \boldsymbol{u}_i, \, \boldsymbol{x}_i\right) = \left\{1 - \mathrm{logit}^{-1}\left(\boldsymbol{u}_i^T \gamma_t + \delta_t H_{i,t-1}\right)\right\} \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}_t + \alpha_t H_{i,t-1}\right), \quad \forall t \in \mathcal{D} \tag{4}$$

the ith day's expected activity count at a time point t. Our choice of the autoregressive term is $H_{i,t-1} = \log(Y_{i,t-1} + 1)$, which gives a convenient interpretation of the coefficient so that $\alpha_t \propto \mathrm{cor}(\log(Y_t), \log(Y_{t-1})) \propto \mathrm{cor}(Y_t, Y_{t-1})$ when $\pi_t \to 0$. Adding one avoids computational problem from a logarithm on zeros. The regression coefficients $\hat{\gamma}_t$, $\hat{\delta}_t$, $\hat{\boldsymbol{\beta}}_t$ and $\hat{\alpha}_t$ are obtained by maximum likelihood estimation with EM Algorithm[27] at each time point t. Due to the conditionality on the past values, we run the model in sequence on $t \in \mathcal{D}$, so that the variables $H_{i,t-1}$ are ready for the model at time t. Equation (4) improves the prediction accuracy for wearing time (see Figure 4) but not the imputation performance for the missing data (see Figure 5). This is for two stated reasons: over-dispersion and the restriction to only include one lag variable. Thus model (4) is not our final model, but it motivates a new proposal in Section 4.3.

4.4 Poisson Log-normal mixture to handle autocorrelation

One of the key features of the model in equation (4) is the autoregressive term, where the model runs successively on minutes and consequently describes the autoregressive relationship in multivariate data. Since $\mathrm{cor}(Y_t, Y_{t-1}) > \mathrm{cor}(Y_t, Y_{t-2})$ the serial effect diminishes over time. This autoregressive assumption fits well for activity data, which has a banding correlation structure as shown in Figure 3(a). In this figure, the activity count at 10:00 a.m. is highly correlated with those at adjacent minutes (bright color) but not necessarily those at 9:30 or 9:00 a.m. (dark color). The question naturally arises as to how many lag or lead variables are sufficient to deal with the autocorrelated activity data. Let K be the lag or lead index from a present time point t; for example, with $K = 2$, we consider the conditional distribution Y_t given $Y_{t-2}, \; Y_{t-1}, \; Y_{t+1}, \; Y_{t+2}$. Note that the model (4) is a regression model, not a stochastic time series model, which means simply including $Y_{t-k}, \; k = 1, \dots, K$ as predictors together will not work

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

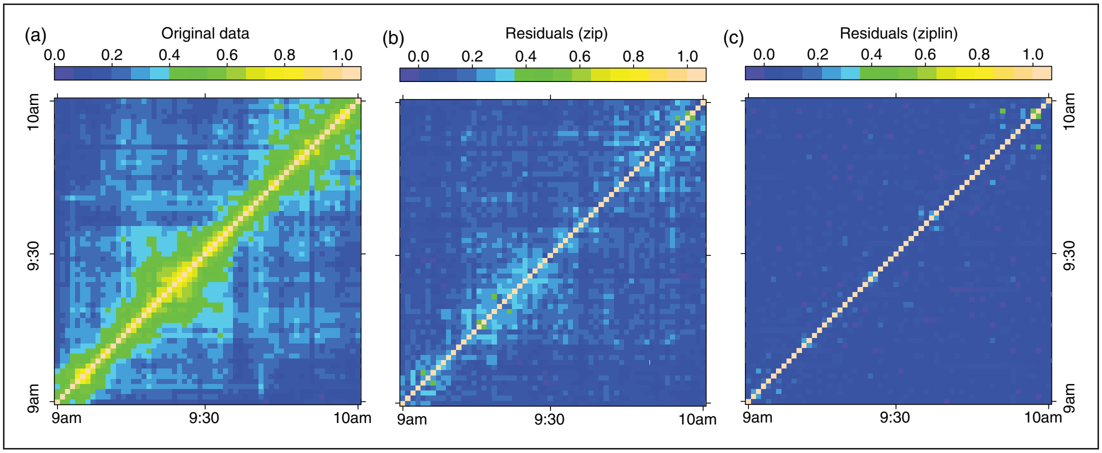

Figure 3. Heatmap of correlation matrix of activity data over 9 a.m. to 10 a.m. Each panel shows a 60 by 60 matrix with the absolute correlation coefficients $|r| < 1$, where the brighter color indicates the higher correlation. In (a), the original activity count data displays the autoregressive correlation structure banding around the diagonals. In (b), the correlation matrix of the residuals from the ZIP model with lag1 term shows that the autoregressive pattern is considerably reduced but not completely. Lastly, (c) displays the correlation matrix of residuals from the ZIPLN model with $K = 3$, showing that the autocorrelation structure disappears.

because of the multicollinearity problem. It is appropriate to frame the model as a conditional multivariate count data in T-dimensional space. Luckily, due to the autoregressive correlation, we only need to include the $2K$ neighboring dimensions.

In what follows, we modify the model (4) assuming the multivariate Poisson Log-normal distribution.[31] Suppose Y follows a multivariate Poisson Log-normal distribution with a dimension of $d = 2K + 1$. This is a mixture of d independent Poisson's and a d-variate Log-normal distribution. With a sample size of N, it can be written as

$$\begin{aligned} Y_i|\lambda_i &\sim \text{Poisson}(\lambda_i) \\ \lambda_i &= \exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta} + e_i\right), \quad \text{for } i \in \mathcal{W} \end{aligned} \tag{5}$$

where $e_i \sim N_d(0, \Sigma)$ with Σ denoting a $d \times d$ variance covariance matrix, i.e., $\lambda_i \sim LN_d(\boldsymbol{x}_i^T\boldsymbol{\beta}, \Sigma)$.

This assumption modifies the model to be

$$\mathrm{E}(Y_{i,t}|\boldsymbol{u}_i, \boldsymbol{x}_i) = (1 - \pi_{it})\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}_t + e_{it}\right), \quad \forall t \in \mathcal{D} \tag{6}$$

For simplicity, we treat $(1 - \pi_{it})$ as a constant due to its role as a weight for the overall expectation which is minimized when $\pi_{it} \to 0$ or $\lambda_{it} \to \infty$. Let $Z_t = \log(Y_t) - \boldsymbol{x}_i^T\boldsymbol{\beta}_t$, and $\boldsymbol{Z} = (Z_{t-K}, \ldots, Z_{t-1}, Z_{t+1}, \ldots, Z_{t+K})^T$, a set of K lag and K lead variables of Y_t. Thus $\boldsymbol{Z} \sim N_{2K}(0, \Sigma_{zz})$. Also denote $\Sigma_{yz} = \mathrm{Cov}(Z_t, \boldsymbol{Z})$. By the property of normal conditional distribution, the imputation model given K lag and K lead variables is expressed by

$$E\left(Y_{i,t}|\boldsymbol{Z}_i, \boldsymbol{u}_i, \boldsymbol{x}_i\right) \propto \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_t + \Sigma_{yz}\Sigma_{zz}^{-1}\boldsymbol{Z}), \qquad \forall t \in \mathcal{D} \tag{7}$$

This model in equation (7) is an update of equation (4) adding both the lag and lead effects.

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

4.5 Test for autocorrelation

We can evaluate these two models (4) and (7) in terms of the amount of autocorrelations remaining in the residuals. It is expected that the residual correlation matrix is close to an identity matrix if the model effectively removes the autocorrelation. Figure 3(b) exhibits the correlation matrix of the residuals from the ZIP regression with one lag term in equation (4), showing that the autoregressive pattern is significantly reduced compared to the original data in (a) but not completely. Panel (c) in this figure displays the covariance matrix of residuals from a mixture of the ZIPLN model in equation (7) with $K = 3$, conditioning on three lag and three lead variables. It is clearly seen that the autocorrelation structure disappears.

In addition to the graphical diagnostics in Figure 3, we can actually test for whether the autocorrelation has been removed after fitting the model. Denote the correlation matrix of residuals $\boldsymbol{R}_{T\times T}$. We perform a test of whether the matrix $\boldsymbol{R}$ is from the uncorrelated data. First, create an $\boldsymbol{I}_0$ matrix, which is a correlation matrix independently drawn from $N\,(0,\ \sigma^2)$ for all $t\ \in\ \mathcal{T}$, setting $\sigma^2 = \hat{\lambda}_t(1 - \hat{\pi}_t)(1 +\ \hat{\pi}_t\hat{\lambda}_t)$ from the marginal ZIP distribution. Second, test the null hypothesis of the equal covariance matrix, $H_0 : \Sigma_R = \Sigma_{I_0}$, meaning the population covariance of $\boldsymbol{R}$ and $\boldsymbol{I}_0$ is equal by applying the Q_2^2 statistic introduced by Srivastava and Yanagihara.[32] The Q_2^2 statistic is computed based on the difference between $tr(\boldsymbol{R})^2/\{tr(\boldsymbol{R})\}^2$ and $tr(\boldsymbol{I}_0)^2/\{tr(\boldsymbol{I}_0)\}^2$ and follows a chi-squared distribution under the null hypothesis, thus a smaller value of Q_2^2 statistic (i.e., large p-value) means greater similarity between the two covariance (or correlation) matrices. This statistic is known to be effective to test the equality of covariance matrix when the dimension is relatively large to the sample size.[32,33] Table 3 summarizes the results from different models indicating that the ZIPLN model with both lag and lead effects has removed the autocorrelation effectively. Furthermore, we can understand the effect of levels of K (how many lag and lead variables are necessary for the imputation model) and find that $K = 1$ is satisfactory at 0.05 level, but $K = 3$ is optimal.

Table 3. The Q_2^2 statistic for the test of the equal correlation matrix of residuals by different models. A smaller Q_2^2 statistic (i.e., large p-value) indicates a great similarity of two correlation matrices, $\boldsymbol{R}$ and $\boldsymbol{I}_0$, meaning the model effectively remove the autocorrelation. The ZIPLN mixture with both lag and lead variables generally perform better with the optimal $K = 3$.

GLM			ZIPLN (lag and lead)		
	stat	*p*-val		stat	*p*-val
Original data	241.54	<0.0001	K=1	3.5217	0.0606
ZIP (L0L1)	15.96	0.0001	K=2	3.0561	0.0804
ZIP (L1L1)	7.47	0.0063	K=3	3.0137	0.0826
ZINB (L1L1)	28.44	<0.0001	K=4	3.0234	0.0821
			K=5	3.0933	0.0786
			K=6	3.1001	0.0783
			K=7	3.1310	0.0768
			K=8	3.1803	0.0745
			K=9	3.1886	0.0742
			K=10	3.2449	0.0716

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

4.6 Prediction accuracy for wearing time

Different models are compared in terms of prediction accuracy. Only wearing time can be evaluated for the prediction accuracy because both the true and predicted values simultaneously exist . If a model predicts well for the wearing time, it should be in the non-wearing time under the MAR assumption. We calculate the prediction accuracy in two ways: the root mean squared errors (RMSE) of counts in equation (8), and the mean area difference (MAD) between the prediction curve and true curve in equation (9). In both measures, a lower value indicates the smaller errors in the prediction, therefore a better predictive model. The first measure is

$$\text{RMSE} = \left\{ \sum_{t \in \mathcal{D}} \sum_{i \in \mathcal{W}} \frac{\left(Y_{it} - \hat{Y}_{it}\right)^2}{|\mathcal{D}||\mathcal{W}|} \right\}^{1/2}, \tag{8}$$

where $|\mathcal{D}|$ indicates the cardinality of a set $\mathcal{D}$, and Y_{it} and $\hat{Y}_{it}$ are the true count and predicted count, respectively, at each i and t. The second measure is

$$\text{MAD} = \sum_{t \in \mathcal{D}} \sum_{i \in \mathcal{W}} \frac{\left|X_i(t) - \hat{X}_i(t)\right|}{|\mathcal{D}||\mathcal{W}|}, \tag{9}$$

where $X_i(t)$ and $\hat{X}_i(t)$ are the smoothing function values at time t fitted from the true count and predicted count, respectively, and the absolute difference is computed. The smoothing function is computed from B-spline with 155 knots.[34]

The results of the prediction accuracy are summarized in Table 4. The ZIPLN mixture model performs better than the standard GLM, especially when both the lag and lead variables are included. Including zero-inflation improves the prediction (ZIP > Poisson). Including a lag 1 variable in the Poisson part improves the prediction (ZIP L0L1 > ZIP). Including a lag1 variable in both logit and Poisson parts improves the prediction (ZIP L1L1 > ZIP L0L1). Zero-inflated Negative Binomial model (ZINB) does not improve the prediction. Furthermore, we compare these models in terms of the Akaike information criterion (AIC) statistic[35] and the non-nested test.[36] These results are summarized in Figure 7 in the Appendix, where the ZINB model performs noticeably better in terms of the AIC because the Negative Binomial model is more

Table 4. Comparison of prediction accuracy for wearing time. The smaller RMSE or MAD indicates the better performance in prediction.

GLM			ZIPLN (lag and lead both)		
	RMSE	MAD		RMSE	MAD
Poisson	543271.0	337.2	K=1	254299.1	118.1
ZIP	543323.1	337.0	K=2	258069.8	129.2
ZIP L0L1	347373.2	160.2	K=3	261339.0	135.6
ZIP L1L1	329017.8	148.4	K=4	263715.5	138.9
ZINB L1L1[a]	383753.2	189.5	K=5	265230.1	140.8

[a] The best GLM model in terms of goodness of fit (AIC).

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

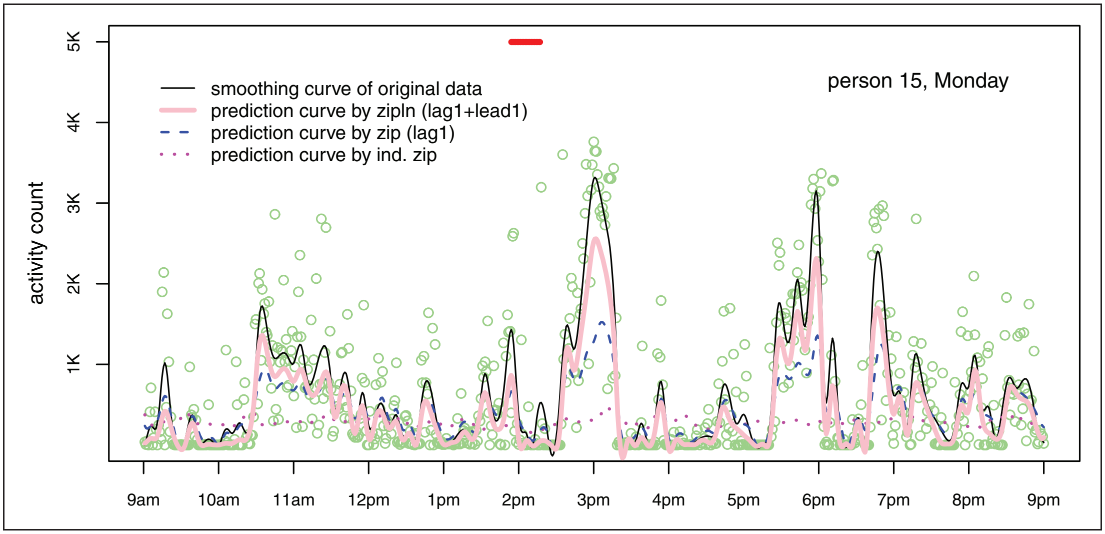

Figure 4. Comparison of prediction curves for wearing time. The closer to the original curve indicates the more accurate prediction.

appropriate for over-disperse data. However, none of these GLM models improve upon any ZIPLN specification since the latter is tailored for both over-dispersion and autocorrelation by including the possible impact of both lagged and leading variables.

Once the minute-successive model is completed, we can display the predicted results for all times in each day, and it allows us to present and compare the prediction curves among methods as shown in Figure 4. In this figure, the prediction curves by different methods are compared to the original curve of the true data points, which is the thin black line. The closer to the original curves indicates the better prediction performance. The difference of the areas between the original curve and other prediction curves means the MAD measure in equation (9). The prediction from the ZIPLN, which is presented by the thick pink line, is the closest to the true curve than any other method. Note that the missing interval is marked with the red line on the top, which is excluded for the computation of prediction accuracy at this time. We will evaluate the imputation accuracy for the missing interval through a simulation in Appendix 1.1.

5 Imputation for missing activity counts

Multiple imputation (MI) is widely accepted as the optimal method for imputing MCAR and MAR missing data ever since Rubin's original work.[37] The basic idea is to create multiple complete datasets by drawing samples from the posterior distribution for each missing value conditional on the observed values, then run multiple parallel models and finally combine the results for one model summary accounting for average within-model variance as well as between-model variance. It is attractive in the sense that the method stresses the uncertainty that could be caused by the missing values instead of estimating the "best" value, which can never be correct. Many extensions and specialized versions have been developed subsequently,[17,38–40] but the core idea of creating replicate full datasets and then combining is preserved.

Our extension to this process applies the fully conditional specification (FSC),[17,39] also known as *multiple imputation with chained equations* (MICE), which is implemented with an open-source

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

software R package.[41] The MI process by FSC is summarized in three steps: (1) model specification, (2) imputation draws from the specified model, and (3) repeated iteration. In our version of this well-known process, we extensively discussed the step (1) in Section 4, and the details for the steps (2) and (3) are provided in the next section.

5.1 Parametric imputation

Parametric imputation is powerful when there is an explicit model that is known or safely assumed. Let Y be an incomplete data vector with missing values and denote $Y^{\text{obs}} = \{Y_i,\ i \in \mathcal{W}\}$ and $Y^{\text{mis}} = \{Y_j,\ _j \in \mathcal{W}^c\}$. For normal data, the imputed value $\hat{Y}_j$ is a draw from $N(X\beta,\ \sigma^2)$. In other words

$$\hat{Y}_j = \quad X\dot{\beta} + \dot{e}$$

where $\dot{\beta}$ is a draw from $N(\beta,\ \sigma^2(X'X)^{-1})$, $\dot{e}$ is a draw from $N(0,\ \sigma^2)$, and the parameters β and σ^2 are estimated from the observed data. With a specified prior distribution, this is called Bayesian imputation under the normal linear model.[37,39] The idea is that the imputation should reflect the parameter uncertainty and prediction errors under the assumption that the missing data distribution is the same as the observed data. The methodological details are well explained in van Buuren[39] and easily accessible through their statistical software.[41] A related procedure for missing outcome variables introduced by Little and Rubin[42] uses the EM Algorithm to cycle between $\hat{Y}_j^{(k)} = X\hat{\beta}^{(k)}$ (E-Step) and $\beta^{(k+1)} = (\widehat{X'X})^{-1}X$ (M-Step) at the kth step for the missing data. Recently, a Bayesian imputation for ZIP and Negative Binomial distribution has been introduced by Kleinke and Reinecke.[43] In this approach, one can draw the missing count values from the estimated probability, $\hat{P}\left(Y_j = h\right), \quad h = 0,\ 1,\ 2, \ldots.$ Unfortunately, our predictive distribution, the ZIPLN, does not have a simplified probability density function that is tractable enough for this approach.[31] Our proposed imputation algorithm, a Bayesian imputation model under the ZIPLN assumptions, is described below in the box.

Imputation Steps under Zero-inflated Poisson Lognormal model

Start with data with initial imputation.

(1) Fit the ZIP model with Y^{obs}at a time point t.
(2) Set $\hat{\boldsymbol{B}} = \left(\hat{\gamma},\ \hat{\boldsymbol{\beta}}\right)^T$ from the coefficients of both logit and poisson models.
(3) Compute the variance-covariance matrix of the coefficients, $\boldsymbol{V} = \text{Cov}(\hat{\boldsymbol{B}})$.
(4) Update the parameters from a posterior distribution $\dot{\boldsymbol{B}} = \hat{\boldsymbol{B}} + \boldsymbol{V}^{-1/2}z$ where $z \sim N(0, 1)$.
(5) Compute $\dot{\pi}_j$ and $\dot{\lambda}_j$ with the updated parameters for Y^{mis}.
$\dot{\pi}_j = \text{logit}^{-1}\left(\boldsymbol{u}^T\dot{\gamma}\right)$
$\dot{\lambda}_j = \exp(\boldsymbol{x}^T\dot{\boldsymbol{\beta}})$
(6) Draw zero imputations based on the $\dot{\pi}_j = \hat{P}(Y_j = 0)$ as follows:
$\hat{Y}_j = 0$ if $\dot{\pi}_j > u_j$, where $u_j \sim \text{unif}(0, 1)$
(7) Draw non-zero imputations based on $\dot{\lambda}_j$ and the log-normal error term.
$\hat{Y}_j = \dot{\lambda}_j \exp(\dot{e})$, where $\dot{e} = \hat{\Sigma}_{yz}\hat{\Sigma}_{zz}^{-1}\boldsymbol{Z}$
with for example setting $K = 1$, $\boldsymbol{Z} = \left[\log\left(Y_{j,\ t-1}\right) - \boldsymbol{x}^T\hat{\boldsymbol{\beta}}_{t-1},\ \log\left(Y_{j,t+1}\right) - \boldsymbol{x}^T\hat{\boldsymbol{\beta}}_{t+1}\right]^T$

This procedure will continue $\forall t \in \mathcal{D}$, which gives a single iteration. Two or three iterations are shown to be sufficient.

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

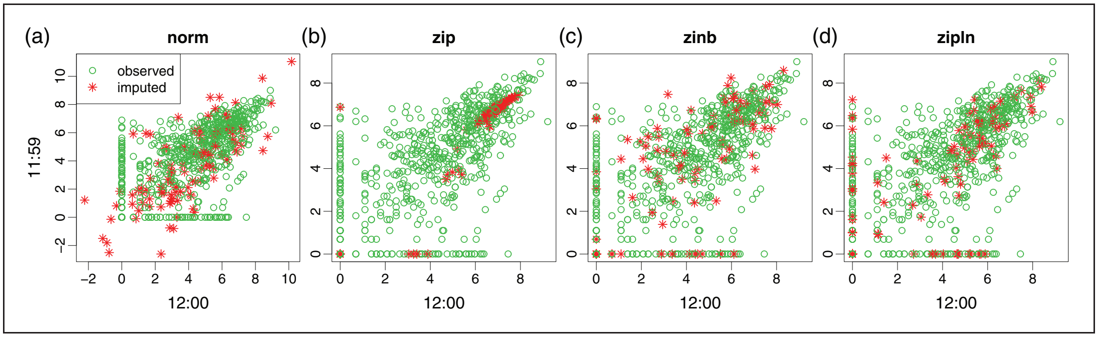

Figure 5. Comparison of parametric imputations at a fixed time point. Each panel displays the scatter plot of activity counts at 12:00 p.m. vs. 11:59 a.m. with log-scale, i.e., log(count+1). The circles are the observed data points, and the stars are the imputed data points.

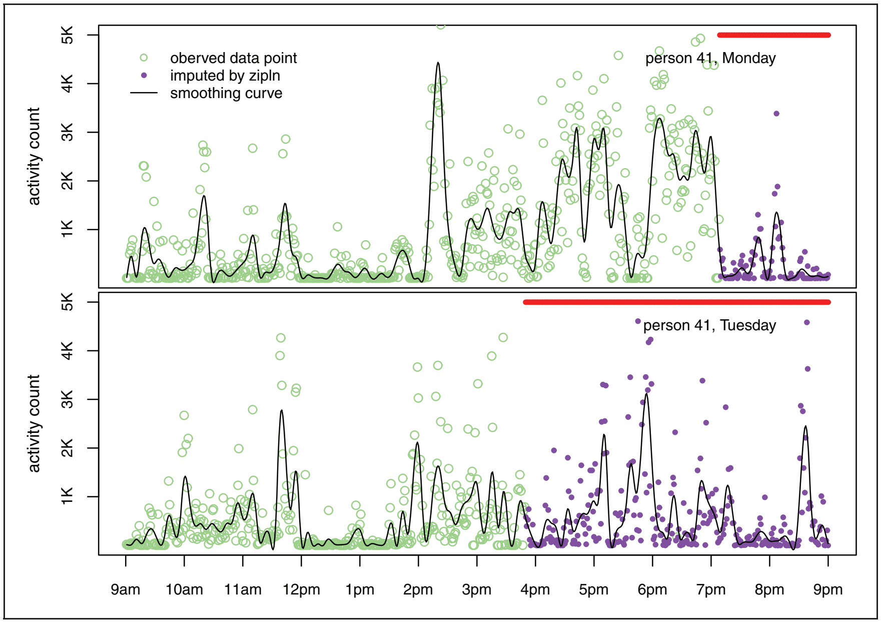

Figure 6. An example of complete data after a Bayesian imputation under the zero-inflated poisson and log-normal mixture model.

Note that the MI involves a doubly iterative process, within a fixed minutes and across the data. A similar approach is found in Nevalainen et al.[44] and Welch et al.,[45] with so called *two-fold* FCS for longitudinal missing data, but these are applied to longer time intervals with only one and one lead.

The imputation difference across different models is given in the panels of Figure 5. Clearly the Normal assumption is inappropriate due to the zero-inflation and non-negative count data as shown in the first panel. The ZIP distribution generates some zero imputations based on the estimated probability of zero, but it fails to handle over-dispersion (second panel). The ZINB distribution shows

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

demonstrably better performance given the over-dispersion problem that the ZIP model is not designed to handle. However, it works less well in the third panel and is computationally intensive. Conversely, the ZIPLN shows much great coalescence with the observed data in the fourth panel of Figure 5.

An example of complete data after a Bayesian imputation under ZIPLN model is displayed in Figure 6.

5.2 Semiparametric imputation

As suggested by Little,[46] we apply the predictive mean matching (PMM) integrated with a Bayesian regression model. The idea is to find the candidate donors among the observed values where each of the missing values is replaced by this process. Rubin[47] originally introduced this method in a simplified form to impute a missing entry Y_j

$$\hat{Y}_j = Y_k$$

where $(\hat{\mu}_j - \hat{\mu}_k)^2 \leq (\hat{\mu}_j - -\hat{\mu}_i)^2$ for all observed values i, $\hat{\mu}_j$ is the predicted mean of Y_j, and Y_k is the observed value that turned out to be the closest candidate for Y_j. A natural extension for the MI process is to find a few number of donors under these circumstances (usually draw three to ten to provide multiple imputed values). Again, Little[46] suggests a refined version using the Bayesian regression specification to reflect uncertainty in estimating parameters. That is, under the standard normal assumption, $\hat{\mu}_j = \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}$ can be replaced by $\tilde{\mu}_j = \boldsymbol{x}_j^T \dot{\boldsymbol{\beta}}$ where $\dot{\boldsymbol{\beta}}$ is a draw from the posterior distribution.

Our second proposal for the activity missing counts is to apply the PMM imputation, as suggested by Little, where the predicted means $\tilde{\mu}_j$, $\hat{\mu}_i$ are provided by the ZIPLN model. Note that the predicted mean values themselves, for any specific models, will not be used as the imputed values, but will be used for the comparison among the profiles. We found that when the ZIPLN model is used, the quality of donors substantialy improved in terms of predictive ability. The simulation study for the imputation accuracy for the missing intervals is provided in Appendix 1.1.

Some literatures describe the PMM as a hot deck imputation.[48,49] One common property is that each missing entry is replaced with an observed value based on the ''similarity'' of the profiles. PMM is one of distance metrics doing this task effectively in that sense. Since a random draw is made among the observed only, it holds the assumption that missing data follow the same distribution as the donors. So the PMM imputation may not be effective when there is not enough data. In terms of the computational time, the parametric method is slightly faster than the semiparmetric PMM procedure that computes all possible pairwise distances.

6 A diagnostic for MAR from the model imputations

There is no general method to test the MAR and the missing completely at random (MCAR) assumptions against not missing at random (NMAR).[25,26] But we can graphically check whether the imputed data from our procedure were reasonably consistent with the observed data under the MAR assumption. Raghunathan and Bondarenko[50] proposed a practical way to check this; that is, if the imputations are reasonable under the MAR assumption, then the Y^{obs} and Y^{mis} should have similar distributions conditional on the propensity score. The propensity scores between 0 and 1, are computed by a logistic regression of a missing indicator vector conditional on some covariates such as age, sex, race, body mass index, and weekday vs. weekend. We found that our proposed imputations for both parametric and semi-parametric processes meet this criterion in general. In doing this, there was no evidence to show violations of the MAR assumption.

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

7 Conclusion

There are methods to handle the correlated multivariate counts data such as multivariate Poisson model[51] or multivariate ZIP model.[52,53] However, such methods are developed for bivariate or trivariate cases rather than the large dimensional data produced by accelerometers. Another critical issue is that the multivariate Poisson (or negative binomial) distribution does not support negative correlation among different discrete random variables,[54] and the probability density function or its other generalized versions are not practical for such applications.[31]

Alternatively, our work uses the multivariate Poisson Log-normal (MVPLN) model, which has also been shown to be useful in many regression applications,[55–57] due to the flexibility that allows for both positive and negative correlation through the normal distribution variance-covariance matrix component. The popularity of MVPLN model is also due to the tractable form of the expectation and variance of this mixture distribution although there does not exist a simplified form of the probability density function.[31]

In addition, we explored the imputation methods for longitudinal time series data[58–61] due to the fact that it is a missing time interval problem. However, the longitudinal missing data typically assume a monotone trend during the missing interval, and this assumption is not suitable for a physical activity time series that has more complex pattern. As shown, the activity data are much noisier and much less smooth than typical longitudinal data and displayed over a very large dimension, usually larger than a sample size. So we also provide prescriptive evidence for future methodological studies of accelerometer data.

The goal of this study is to find an effective method for imputing accelerometer missing data, which we have shown to have very different characteristics than conventional rectangular datasets of $N > T$. Our main adjustment lies in specifying a viable imputation distribution with a mixture distribution of ZIPLN. Using this specification, multiple imputation by chained equations are applied for $N \times T$ accelerometer data matrix ($N < T$). We then demonstrate that under the ZIPLN model, both the parametric and semi-parametric imputations work better than other competing methods in imputation performance. Therefore, this missing data method is a useful addition to the literature for dealing with general multivariate count data with under-dispersion (zero-inflation) and over-dispersion (autocorrelation) problems.

8 Software

To facilitate the practical use of this method, we provide an R package *accelmissing*, which can be incorporated with the existing *mice* and *pscl* R packages.

Funding

This research was supported by National Cancer Institute at the National Institutes of Health (U54 CA155496) through the Transdisciplinary Research on Energetics and Cancer Center at Washinton Univeristy in St. Louis.

References

1. Physical Activity Guidelines Advisory Committee. *Physical activity guidelines advisory committee report, 2008*. Washington, DC: U.S. Department of Health and Human Service, 2008.
2. Durante R and Ainsworth BE. The recall of physical activity: using a cognitive model of the question-answering process. *Med Sci Sports Exer* 1996; **28**: 1282–1291.
3. Jacobs DR Jr, Ainsworth BE, Hartman TJ, et al. A simultaneous evaluation of 10 commonly used physical activity questionnaires. *Med Sci Sports Exer* 1993; **25**: 81–91.
4. Catellier DJ, Hannan PJ, Murray DM, et al. Imputation of missing data when measuring physical activity by accelerometry. *Med Sci Sports Exer* 2005; **37**: S555–S562.
5. Metzger JS, Catellier DJ, Evenson KR, et al. Patterns of objectively measured physical activity in the united states. *Med Sci Sports Exer* 2008; **40**: 630–638.

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

6. Troiano RP, Berrigan D, Dodd KW, et al. Physical activity in the united states measured by accelerometer. *Med Sci Sports Exer* 2008; **40**: 181–188.
7. Evenson KR and Wen F. Performance of the actigraph accelerometer using a national population-based sample of youth and adults. *BMC Res Notes* 2015; **8**: 7.
8. Choi L, Liu Z, Matthews CE, et al. Validation of accelerometer wear and nonwear time classification algorithm. *Med Sci Sports Exer* 2011; **43**: 357–364.
9. Masse LC, Fuemmeler BF, Anderson CB, et al. Accelerometer data reduction: a comparison of four reduction algorithms on select outcome variables. *Med Sci Sports Exer* 2005; **37**: S544–S554.
10. Miller GD, Jakicic JM, Rejeski WJ, et al. Effect of varying accelerometry criteria on physical activity: the look ahead study. *Obesity* 2013; **21**: 32–44.
11. Winkler EA, Gardiner PA, Clark BK, et al. Identifying sedentary time using automated estimates of accelerometer wear time. *Brit J Sports Med* 2012; **46**: 436–442.
12. Trost SG, McIver KL and Pate RR. Conducting accelerometer-based activity assessments in field-based research. *Med Sci Sports Exer* 2005; **37**: S531–S543.
13. Nilsson A, Ekelund U, Yngve A, et al. Assessing physical activity among children with accelerometers using differerent time sampling intervals and placements. *Pediat Exer Sci* 2002; **14**: 87–96.
14. Lee PH. Data imputation for accelerometer-measured physical activity: the combined approach. *Am J Clin Nutr* 2013; **97**: 965–971.
15. Chinapaw MJ, de Niet M, Verloigne M, et al. From sedentary time to sedentary patterns: accelerometer data reduction decisions in youth. *PLoS One* 2014; **9**: e111205.
16. Altenburg TM, de Niet M, Verloigne M, et al. Occurrence and duration of various operational definitions of sedentary bouts and cross-sectional associations with cardiometabolic health indicators: the energy-project. *Prev Med* 2015; **71**: 101–106.
17. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res* 2007; **16**: 219–242.
18. Stevens J, Murray DM, Catellier DJ, et al. Design of the trial of activity in adolescent girls (TAAG). *Contemp Clin Trials* 2005; **26**: 223–233.
19. Cohen DA, Ashwood JS, Scott MM, et al. Public parks and physical activity among adolescent girls. *Pediat* 2006; **118**: e1381–e1389.
20. Webber LS, Catellier DJ, Lytle LA, et al. Promoting physical activity in middle school girls: trial of activity for adolescent girls. *Am J Prev Med* 2008; **34**: 173–184.
21. Pate RR, Stevens J, Webber LS, et al. Age-related change in physical activity in adolescent girls. *J Adolesc Health* 2009; **44**: 275–282.
22. Evenson KR. Towards an understanding of change in physical activity from pregnancy through postpartum. *Psychol Sport Exer* 2011; **12**: 36–45.
23. Cradock AL, Wiecha JL, Peterson KE, et al. Youth recall and tritrac accelerometer estimates of physical activity levels. *Med Sci Sports Exer* 2004; **36**: 525–532.
24. Morris JS, Arroyo C, Coull BA, et al. Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: a case study. *J Am Stat Assoc* 2006; **101**: 1352–1364.
25. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc* 1996; **91**: 473–489.
26. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**: 581–590.
27. Lambert D. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**: 1–14.
28. Cohen AC. *Estimation in mixtures of discrete distribution*. reprint ed.: Statistical Pub. Society, 1963.
29. Yip P. Inference about the mean of a poisson distribution in the presence of a nuisance parameter. *Aust J Stat* 1988; **30**: 299–306.
30. Mullahy J. Specification and testing of some modified count data models. *J Econometric* 1986; **33**: 341–365.
31. Aitchison J and Ho CH. The multivariate poisson-log normal-distribution. *Biometrika* 1989; **76**: 643–653.
32. Srivastava MS and Yanagihara H. Testing the equality of several covariance matrices with fewer observations than the dimension. *J Multivar Anal* 2010; **101**: 1319–1329.
33. Lee JA, Dobbin KK and Ahn J. Covariance adjustment for batch effect in gene expression data. *Stat Med* 2014; **33**: 2681–2695.
34. Ramsay JO, Hooker G and Graves S. *Functional data analysis with R and MATLAB*. New York: Springer, 2009.
35. Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Control* 1974; **19**: 716–723.
36. Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 1989; **57**: 307–333.
37. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: John Wiley&Sons, 1987.
38. Schafer JL. *Analysis of incomplete multivariate data*. London: Campman&Hall/CRC, 1997.
39. van Buuren S. *Flexible imputation of missing data*. Boca Raton, FL: Champman&Hall/CRC, 2012.
40. Gelman A. Parameterization and bayesian modeling. *J Am Stat Assoc* 2004; **99**: 537–545.
41. van Buuren S and Groothuis-Oudshoorn K. mice: Multivariate imputations by chained equations in R. *J Stat Softw* 2011; **45**: 1–67.
42. Little RJA and Rubin DB. *Statistical analysis with missing data*, 2nd ed. New York: John Wiley & Sons, 2002.
43. Kleinke K and Reinecke J. Multiple imputation of incomplete zero-inflated count data. *Stat Neerl* 2013; **67**: 311–336.
44. Nevalainen J, Kenward MG and Virtanen SM. Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification. *Stat Med* 2009; **28**: 3657–3669.
45. Welch CA, Petersen I, Bartlett JW, et al. Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data. *Stat Med* 2014; **33**: 3725–3737.
46. Little RJA. Missing-data adjustments in large surveys. *J Bus Econ Stat* 1988; **6**: 287–296.
47. Rubin DB. Statistical matching using file concatenation with adjusted weights and multiple imputations. *J Bus Econ Stat* 1986; **4**: 87–94.
48. Andridge RR and Little RJA. A review of hot deck imputation for survey non-response. *Int Stat Rev* 2010; **78**: 40–64.
49. Cranmer SJ and Gill J. We have to be discrete about this: A non-parametric imputation technique for missing categorical data. *Brit J Polit Sci* 2013; **43**: 425–449.
50. Raghunathan T and Bondarenko I. *Diagnostics for multiple imputations*. Available at SSRN: http://ssrn.com/abstract=1031750 (accessed 21 November 2007).
51. Marshall AW and Olkin I. A family of bivariate distributions generated by the bivariate bernoulli distribution. *J Am Stat Assoc* 1985; **80**: 332–338.
52. Li CS, Lu JC and Park JH. Multivariate zero-inflated poisson models and their applications. *Technometrics* 1999; **41**: 29–38.
53. Walhin JF. Bivariate zip models. *Biometrical J* 2001; **43**: 147–160.

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

54. Holgate P. Estimation for bivariate poisson distribution. *Biometrika* 1964; **51**: 241–245.
55. Chib S and Winkelmann R. Markov chain monte carlo analysis of correlated count data. *J Bus Econ Stat* 2001; **19**: 428–435.
56. Ma J, Kockelman KM and Damien P. A multivariate poisson-lognormal regression model for prediction of crash counts by severity, using bayesian methods. *Accid Anal Prev* 2008; **40**: 964–975.
57. El-Basyouny K and Sayed T. Collision prediction models using multivariate poisson-lognormal regression. *Acciden Anal Prev* 2009; **41**: 820–828.
58. Dunsmuir W and Robinson PM. Estimation of time-series models in the presence of missing data. *J Am Stat Assoc* 1981; **76**: 560–568.
59. Jones RH. Maximum-likelihood fitting of ARMA models to time-series with missing observations. *Technometrics* 1980; **22**: 389–395.
60. Shumway RH and Stoffer DS. An approach to time series smoothing and forecasting using the em algorithm. *J Time Ser Anal* 1982; **3**: 253–264.
61. Honaker J and King G. What to do about missing values in time-series cross-section data. *Am J Polit Sci* 2010; **54**: 561–581.

Appendix

A1.1. Imputation accuracy for missing intervals: A simulation study

Imputation by predictive mean matching (PMM) relies on a single summary statistic or the distances computed by these statistics, instead of the distributional parameters. Because of

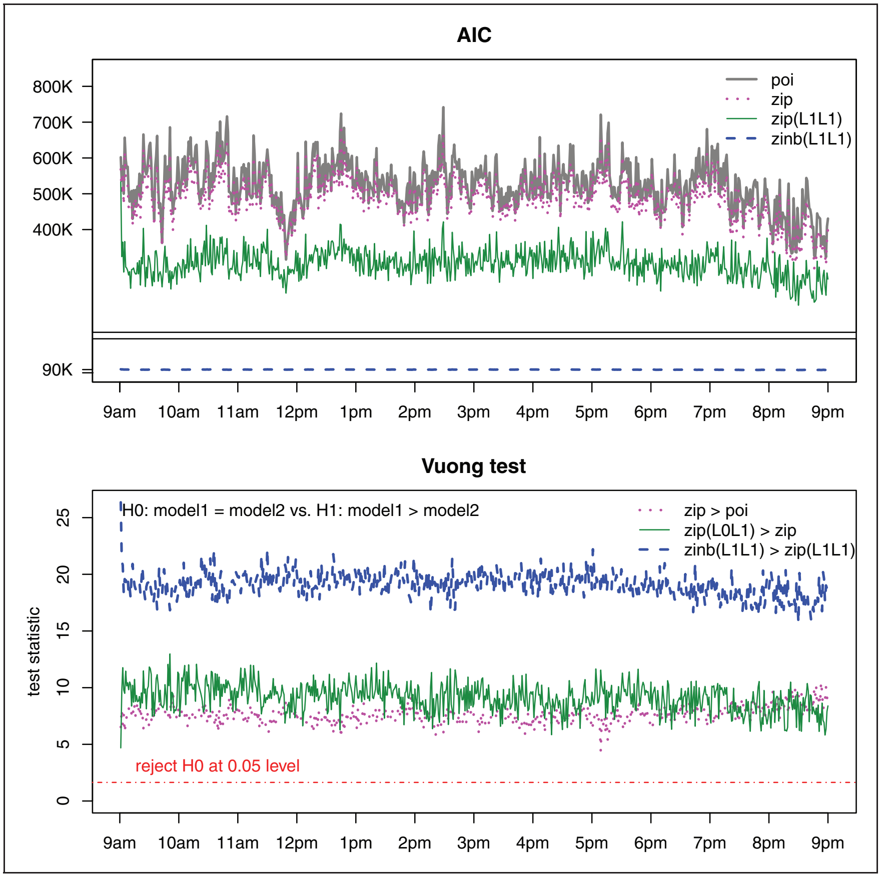

Figure 7. Model comparison by Akaike information criterion (AIC) and Vuong test. The smaller AIC indicates the better model. The high vuong statistic above the dashed-dotted line at 1.64 means the null hypothesis of equivalent models is rejected at 0.05 significance level, i.e., the model 1 is better than the model 2.

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

such simplicity and a lack of strict assumptions, PMM imputation is robust against misspecification of the imputation model. Figure 8 also supports this fact by showing reasonable imputations regardless of model specification in PMM imputations. However, there are some obvious differences in terms of the donors' predictive ability, which may depend on how effectively the specified model, as a distance metric, captures the similarity or dissimilarity among the daily profiles.

Imputation is a form of prediction, so it is expected that a set of donors should contain predictive information for a target missing value if the choice of donors were reasonable. In order to evaluate

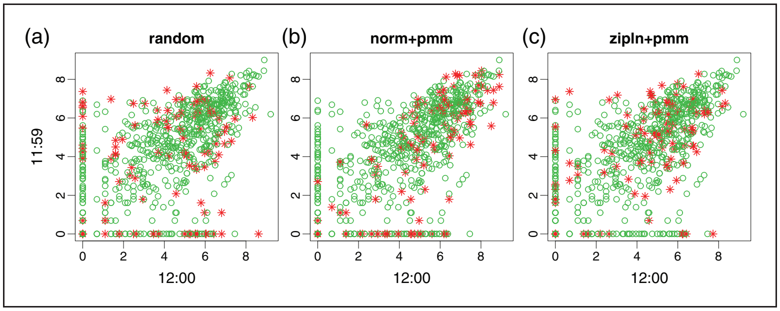

Figure 8. Comparison of PMM imputation at a fixed time point. The PMM imputations produce reasonable imputations regardless of model specification.

Table 5. Imputaion accuracy of simulated missing data. The average of *D* donors are compared to the true value by RMSE and MAD. The smaller value indicates the better performance. Using ZIPLN model outperforms other methods.

Imputation method	RMSE	MAD
Mean	557137.4	331.5
Random sample (D=1)	1193682.3	375.7
Random sample (D=3)	711333.3	347.0
Random sample (D=5)	665840.1	339.8
Random sample (D=10)	612203.1	333.7
NORM+PMM (D=1)	1098393.7	421.3
NORM+PMM (D=3)	916131.1	454.8
NORM+PMM (D=5)	833960.6	454.0
NORM+PMM (D=10)	769614.4	458.1
ZIP+PMM (D=1)	1080756.2	426.2
ZIP+PMM (D=3)	834745.1	432.2
ZIP+PMM (D=5)	760667.5	432.4
ZIP+PMM (D=10)	697649.9	433.0
ZIPLN+PMM (D=1)	505524.6	141.9
ZIPLN+PMM (D=3)	348106.6	113.9
ZIPLN+PMM (D=5)	319995.3	108.3
ZIPLN+PMM (D=10)	301809.5	106.4
ZIP (m=5)	691990.4	476.8
ZINB (m=5)	696653.4	357.0
ZIPLN (m=5)	308985.3	145.1

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016

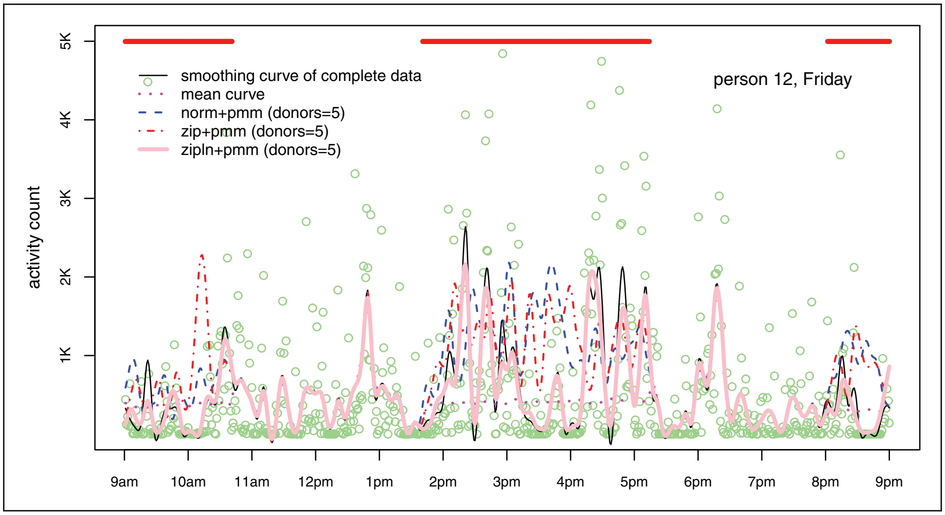

Figure 9. Comparison of imputation curve by different PMM methods. Five donors selected by each method are averaged to impute the artificial missing interval (red line on the top). It is clear that the imputation from the ZIPLN model performs superior since its imputation is closest to the true data, implying that the selected donors by this method contain better predictive information than other methods.

the imputation accuracy, we need a simulation study according to the following steps. First, randomly generate the missing intervals of 20−180 min length. Second, produce imputations under a specific model. Third, compare the true vs. imputed value. Imputation accuracy is computed by RMSE and MAD, as done in Section 4.6. A single imputed value, averaged from D donors, is compared to the true value.

For comparison, we impute the simulated missing data with four methods: (1) draw a random sample from all observed values, (2) select donors based on the smallest distances $d(\tilde{\mu}_j, \hat{\mu}_i)$ under the normal linear model (NORM+PMM), (3) select donors based on the smallest distances $d(\tilde{\mu}_j, \hat{\mu}_i)$ under the zero-inflated Poisson model (ZIP+PMM), and (4) select donors based on the smallest distances $d(\tilde{\mu}_j, \hat{\mu}_i)$ under the zero-inflated Poisson Lognormal model (ZIPLN+PMM). The results are summarized in Table 5. Note that the random sample imputation does not have any predictive information. The PPM imputations by NORM or ZIP do not excel the performance of the random imputation or even worse. The ZIPLN model outperforms the other methods. We also compare these to the parametric imputation methods in which we use the average of five multiple datasets ($m = 5$). Here, the ZIPLN performs the best as well. Graphical comparison is also provided in Figure 9.

Downloaded from smm.sagepub.com at WASHINGTON UNIV SCHL OF MED on November 28, 2016