Missing Data Imputation for Time Series Accelerometry Data

Deji Suolang, Kaidar Nurumov, Yongchao Ma

Introduction

The wearable device has emerged as an important mean of assessing human behaviors and served to define outcome measures in health observational and experimental studies. However, instances such as non-wearing of the device or partial wear can lead to an underestimation of total activity. Moreover, there might be potential disparities in activity levels between individuals with complete and incomplete or even no data, thereby rendering the estimated summary statistics from complete days susceptible to selection bias. This study focuses on the measurement of sedentary time during the day (8am–10pm).

This research aims to address this challenge by pioneering and assessing various imputation methods for the missing time series data to mitigate underreporting bias, a crucial factor influencing the quality of wearable tracking data. This effort not only carries immediate relevance for ongoing investigations but also offers enduring value for future researchers to contemplate, extending beyond the confines of the present analysis result. All materials to reproduce this research are available on GitHub.

Data and Methods

Data

We use data from the Physical Activity and Transit Survey (PAT). The survey asked adults in residential households in New York about their physical activity at work, home, commuting, and recreation. Among the 3811 respondents, 679 of them consented to wear accelerometer devices during all waking hours for one week, and had valid data (>4 days, >10 hrs per day). We created a concatenated dataset consisting of self-reported surveys and minute-level accelerometer data. The daytime raw accelerometer data has 3,992,520 observations, and the missing rate is about 20%.

Methods

As the key estimate of interest is a sedentary minute in a week, the target variable in the imputation can be obtained in two different ways: 1) imputing the missing activity count, and then using the thresholds in the literature to define whether a minute is sedentary. We use the Freedson (1998) cut-off point of [0-100) count-per-minute to define sedentary minutes in the accelerometer data. 2) As an alternative, we can directly impute a binary indicator of whether the missing minute is a sedentary minute. There are several similar studies in the previous literature attempting to tackle the missing data problem in the time series accelerometry data. Three common methods are linear interpolation, zero-inflated Poisson models, and ARIMA models (R packages: 'accelmissing', 'imputeTS') (Lee et al, 2019; Moritz & Bartz-Beielstein, 2017).

In this research, we first analyze the relationship between missingness and minute-level variables, as well as missingness and survey variables reflecting individual's characteristics. Then, based on single and multiple imputation approaches we test four different methods: 1) Single imputation using traditional time series methods 2) Multiple imputations for imputing minute-level missing activity count, using time-varying variables and survey variables as predictors; 3) Multiple imputations for imputing the binary indicator for a sedentary minute; 4) XGboost ML algorithm to impute activity count and sedentary minute, treating both as cross-sectional. Finally, we compare our methods against the complete case approach and survey self-reports using sum of sedentary minutes in a week across each individual.

This study answers the following research questions: (1). *How does the imputation result in different key estimates compared to the complete cases method?* (2). *How do the accelerometry-based estimates compare to the self-reports in the survey?* (3). *How do different modeling approaches perform?* (4). *Does imputing activity count and determining sedentary status, as opposed to imputing sedentary status, enhance the accuracy and reliability of activity data imputation?*

Computational challenge

Sensory digital traces are a new type of big data with a complex structure and large size. Modeling the minute level can be challenging due to the non-convergence because there are about 5,880 minute-level observations per participant and the data from 679 participants resulted in about 4 million minute-level records. The computational intensity of these methods will be a focal point of our resource planning and execution strategy. Where possible we used a random, stratified sample of observations to initially test the model performance. Where necessary, we conducted a code optimization for the parallel running. The analyses are submitted as a slurm job on the Great Lakes High-Performance Computer, which provides us the memory and storage capability beyond our local machine.

Results

Descriptive Analysis of Missing Data

Each participant has 5,880 minute observations (14 hours per day for 7 days). Of the 679 participants with accelerometer data, 679 participants have missing minute observations. The following figure on the left shows the distribution of missing minute observations. On average, each participant has 1,153 missing minute observations (indicated by the red dashed line). In total, 781,940 minute observations are missing. It is computationally challenging to impute missing minute observations.

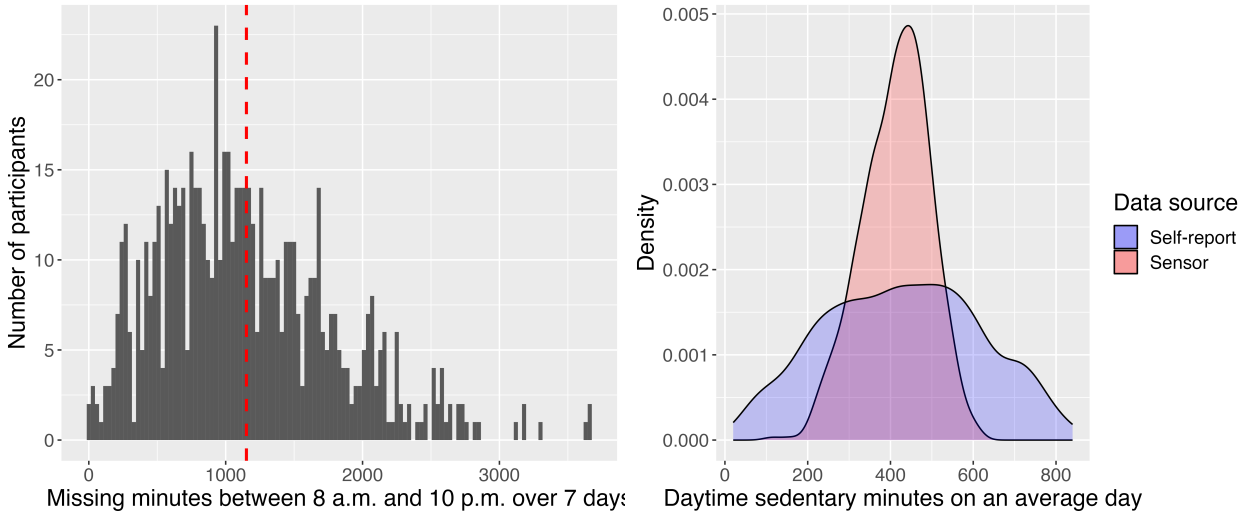
The measurement of interest is the daytime sedentary minutes. Besides the measure derived from the accelerometer data, a survey question asks about the minutes of sedentary activity on an average day. The following figure on the right shows the densities of two variables measuring the daytime sedentary minutes on a day. Point estimates of average sedentary minutes converge between the two measures but the variance of the self-reported survey variable is greater than the variance of the accelerometer variable. It suggests that participants may under/over-report sedentary activities due to recall bias. As accelerometer sensors collect more accurate data, it may be tempting to obtain complete accelerometer data.

Complete Case Approach

The complete cases approach is a straightforward method for handling missing data, which analyzes only those observations for which there are no missing values. We obtained an average of 2,894 minutes (SE 570.89) of daytime sedentary time in a week.

Linear Interpolation

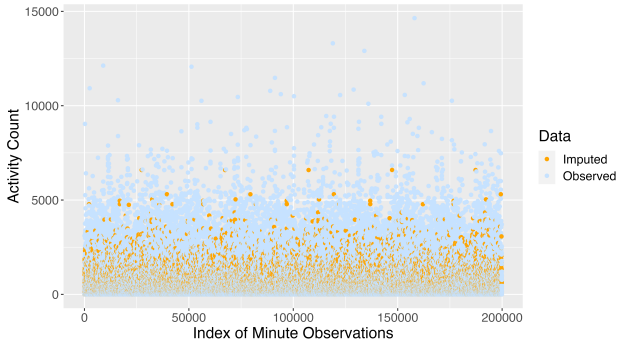
Linear interpolation is a method of estimating values between two known values in a dataset. It assumes a linear relationship between the known values, and it calculates intermediate values based on this assumption. We found the average daytime sedentary minutes based on linear interpolation is 2,991 minutes (SE 574.16).



Multiple Imputation

Impute activity count using minute-level and individual level variables We first conducted the multiple imputation for minute-level activity count. To stabilize the variable, we used the square root transformation and back-transformed after the imputation. We defined the target variable activity count as a non-negative continuous variable. The imputation involved 10 iterations and generated m=5 datasets. Multiple imputations capture the variability between imputations.

The predictors in the imputation model include 1) Time-varying variable lagged activity count (square-root transformed), which is the last activity count observed before the one at the missing time point; 2) Hour's position in the day; 3)Individual-level characteristics such as age, gender, race, activity limitation, income/poverty ratio, BMI range. They are the variables proven to be relevant to the sedentary behavior in the literature, and a result of backward covariate selection and variable importance test from the XGboost Tree in the below.



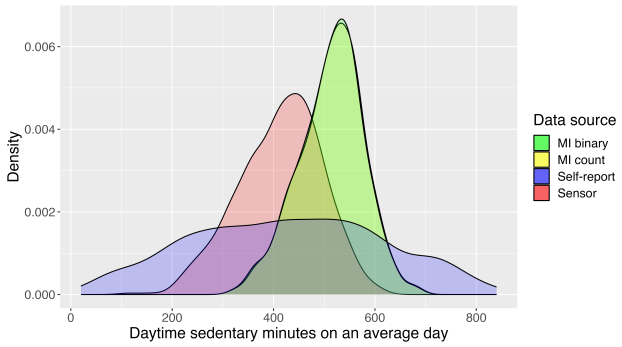
The observed activity count has a mean of 335, while the mean of the imputed value is 581. The imputation is based on the full data. However, given there are millions of data points, we randomly sampled 10% of them for the diagnostic plot visualizing the observed vs. imputed values diagnostic plot. We observed their distributions are similar.

After we had the complete dataset of 5, we pooled the estimate by taking the average activity count from 5 datasets for each individual. Similar to what we have done earlier, we then use the activity count threshold to determine whether a minute is a sedentary minute. Multiply imputing approach activity count provided us with a mean of 3,234 minutes (SE 591) of sedentary minutes in a week.

Impute activity count and binary sedentary status using minute-level variables We used multiple imputation to fill the missing minute observations given the relationship between the activity count/binary

indicator of sedentary status and three fully-observed minute-level variables (timestamp, day of week, and hour of day). We found that the missingness is not equally distributed across different values of the timestamp, day of week, and hour of day. Specifically, missingness is more likely to happen in the morning and evening, as well as on Saturday and Sunday.

We obtained five datasets with imputed activity counts and five datasets with imputed binary sedentary status. To improve the computational efficiency, five cores were used to generate the imputed datasets in parallel. Predictive mean matching was used to impute the activity count and logistic regression is used to impute the binary indicator of sedentary status. The following figure shows that the distribution of sedentary minutes derived from the imputed activity counts and binary indicators of sedentary status are similar to each other.



XGBoost

In addition to multiple imputation we used a machine learning approach where unlike the previous approaches, we used complete case data to train (70%) and test (30%) our models and then filled in missing values based with the trained model. To train our model we used XGBoost, an efficient regularized boosting algorithm that uses a sequential ensemble of trees to improve model performance. The algorithm can be used with binary, categorical as well as continuous variables. Overall, to train the model we selected 60 complete case covariates (features) available for the datasets with missing and non-missing target variables. The model was trained using a set of hyperparamaters specific to xgboost and 16 threads for faster computation. Since we have a complex data structure we increased number of boosted trees (nrounds) to 10000. The trained models were used to predict the continuous count and binary activity for the test as well as the data with missing sed_min and count activity labels. For the test dataset, we evaluated the model performance using balanced accuracy (binary_count) and R-squared (count data) sensitivity, and specificity calculated based on the confusion matrix of predicted and observed sedentary behavior cases.

The results of model with count activity produced R-square value of 0.47 whereas for the binary data, the overall accuracy was 0.72 with sensitivity of 0.87 and specificity of 0.50. Top 3 most important features for both models included individual level variables such as minuteid, weartime and hourid. These results tell us that the model can predict correctly on average 9 out of 10 cases with sed_min=1 and 5 out of 10 cases with sed_min=0. The balanced accuracy is 0.69. Using the trained models we first imputed missing values for activity count with the subsequent calculation of sedentary minutes (mean - 2,746 and SE - 670) after this step we imputed sedentary minutes directly (mean - 2,720 and SE - 550).

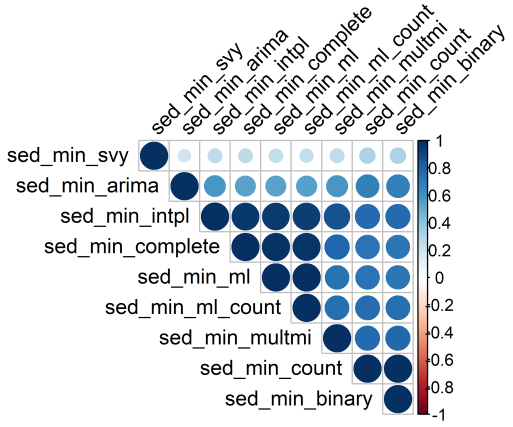
Autoregressive Integrated Moving Average and Kalman Smoothing

While the machine learning approach provided a convenient way of filling NAs using ensemble training we did not account for the time series nature of accelerometer count activity data. Thus, we used the ARIMA model for the univariate time series count together with Kalman smoothing (KS). While the ARIMA model uses a lagged moving average, KS can account for noisy measurement by incorporating information from past and future observations and hence improve the imputation of time-series data. Noteworthy, for a sequence of missing values using the transition equation KS can make a best guess without the data, hence the imputed NAs can take any value including the negative. For our imputed data, we replaced all negative imputations

with zeros. The results imputed for each individual separately, show that the imputed mean value of sedentary minutes in a week across individuals is 3,256 and SE is 681

Comparison of Estimates from Different Appraoches

The complete case approach deleted all missing values and therefore at the risk of underestimation. Among the time series imputation results, we found that linear interpolation produced the smallest estimate and standard error. As interpolation uses the mean value before and after the missing time point, a smaller variation is expected. With multiple imputation approaches, imputing binary indicators produced smaller estimates and smaller standard errors. We observed correlations between two multiple imputation methods are high. The machine learning results produced the smallest mean and SE among all approaches, however based on the balanced accuracy, specificity and R-square metrics we can conclude that these results require careful interpretation. We found the accelerometer-based estimate and self-reports have low correlations around 0.2. It is not surprising as such weak correlations are repeatedly confirmed by the literature. We conclude these two measures have different properties.



Discussion

This study added additional methods to the existing literature for handling missing data in the time series accelerometer data. We demonstrated the estimates based on different methods and provided recommendations for future research. The activity count as a proxy of acceleration wearable device detects is not a counting process of the random independent events, therefore Poisson or negative binomial distributions that usually apply to count data do not apply in this case. We treated it as a continuous variable, but this is under debate. The lack of predictive power may be due to the lack of suitable distribution for such data. In addition, there are several limitations: (1) Due to very bad training results we did not use the LSTM model (2) We did not use multilevel models for the imputation, due to low, but significant random variance as well as convergence issues for the full sample. (3) Since we used different imputation approaches our imputation models are not always directly comparable. Future research can build on these limitations and further investigate ways to efficiently impute accelerometer data. The takeaways from this study: 1) Understanding missing patterns is the most critical first step; 2) as there is no any widely-accepted statistical distribution for such accelerometer-based activity count, predicting binary activity status instead of raw count might be a more reliable option; 3) The use of ML for imputation is promising, we managed to improve the xgboost performance by dramatically increasing the number of trees, first from 500 to 1000 and then to 10000. Our attempt can be further improved by training the model on a more comprehensive grid of hyperparameters.

Author Contribution Statement

Deji Suolang is responsible for multiple imputations, Kaidar Nurumov is responsible for training machine learning models and time-series imputation using ARIMA, Yongchao Ma is responsible for descriptive analysis, a compilation of the final report, and maintenance of the GitHub repository. All team members contributed to the study design, data management, and report writing.