**Predicting sleep quality using accelerometer data from wearable devices**

Casandra Philipson

Foundations in Data Science

Springboard

March 2016

Mentor: Marko Mitic

**1. Introduction**

**1.1 Significance and Overview**

Wearable lifestyle tracking devices have rapidly evolved into a billion-dollar market facilitating real-time monitoring and ample opportunity to manage health in a personalized manner. Moreover, constant quantification of physical activity, heart rate, and sleep quality offers substantial insight for preventative health care. In many cases, big data obtained from wearable devices are simply x, y, and z coordinates with a timestamp indicating the collecting time. Predictive algorithms turn the millions of actigraphic measurements into meaningful information via statistics and machine learning. Although most fitness websites and apps accurately classify activity levels (e.g. laying, walking, running) sleep quality is often overestimated, especially for individuals suffering from sleep disorders. Thus continued optimization for sleep cycle tracking is valuable. This project uses data obtained from wrist-worn accelerometer trackers on healthy subjects and patients suffering from sleep apnea syndrome and validates the need to improve predictions for quality of sleep.

**1.2 Studying Sleep**

*Sleep cycle –* In healthy individuals, a typical sleep cycle alternates between three non-REM (NREM) phases followed by REM (Table 1 and Figure 1). The average sleep cycle lasts around 120 minutes.

**Table 1. Sleep phase description**

|  |  |  |
| --- | --- | --- |
| Data set Value | Phase | Length/episode |
| 3 | NREM 1 | Seconds to minutes |
| 2 | NREM 2 | 10 to 15 minutes |
| 1 | NREM 3 | 20 to 40 minutes |
| 5 | REM | 1 to 5 minutes (or more) |
| 6 | Awake | NA |
| 7 | Movement | NA |

*Sleep disorders –* Sleep deprivation has a detrimental impact on human life. An estimated 30% of adults suffer from sleep disorders. In the short term, lack of sleep can cause memory loss, decreased performance, reduced alertness, and increased stress. If sleep deprivation persists long term, perhaps as a result of undiagnosed sleep disorders, an individual is significantly more prone to heart failure, stroke, obesity, depression, and mental impairment. Some sleep disorders include sleep apnea, insomnia, and restless leg syndrome. Additionally, Posttraumatic Stress Disorder (PTSD) is associated with significant sleep disturbance. In this study patients suffer from sleep apnea, a disorder characterized by shallow or paused breathing during sleep leading to a buildup of carbon dioxide that subsequently wakes the brain.

*Sleep Analysis –* Gold standard sleep studies in medicine use polysomnography (PSG) to quantify brain waves, brain oxygen, heart rate, breathing, in addition to eye and leg movements.Brain function (EEG), eye movement (EOG), and muscle activity (EMG) patterns obtained from a PSG effectively capture REM and NREM sleep phases. Due to affordability, ease, and accessibility for patients,actigraphic measurements have been employed as an alternative method for studying sleep-wake rhythms. However, improved sleep estimation algorithms are desired for widespread acceptance in the medical community.

**1.3 Wearable Devices with Accelerometers for Sleep Tracking**

People are able to track sleep using a variety of gadgets spanning from a ring sized wellness wearable to smart pillows. Smartphone and iWatch apps such as Sleep Cycle, Sleepbot, and Sleep Time analyze sleep and go as far as making predictions to wake the user during the lightest sleep phase to reduce morning fog. Fitness watches produced by FitBit, Jawbone, and Nike+ contain actigraph sensors for sleep and activity tracking and have the advantage of measured heart rate (models and data closed-sourced). GENEActiv and Actiwatch are two actigraph devices that are clinically evaluated and release raw data to users.

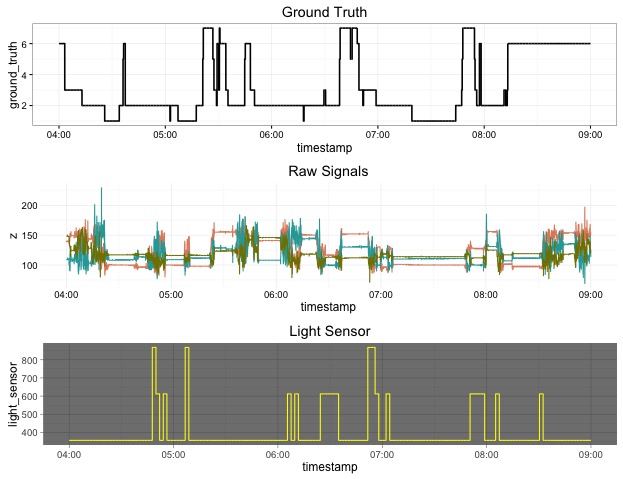
**2. Actigraphic Sleep Data**

**2.1 Publically available data**

The data set used in this project is open source and available for download here: <http://www.ess.tu-darmstadt.de/ichi2014>). Raw data was obtained from 42 patients (aged between 28 and 86 years) using a 100Hz wrist-worn data-logging device. The majority of patients suffer from a sleep disorder (insomnia, narcolepsy, sleep apnea syndrome, or restless leg syndrome). Diagnoses for each patient are recorded. Healthy control subjects are also included in the study.

**2.2 Contents of the data set:**

* Unix timestamp
* Run length encoding
* 3D accelerometer values for *x, y,* and *z*
* Light sensor values
* Ground truth sleep phases determined by polysomnography
  + 3 = NREM1, 2 = NREM2, 1 = NREM3, 5 = REM, 6 = Awake, 7 = Movement

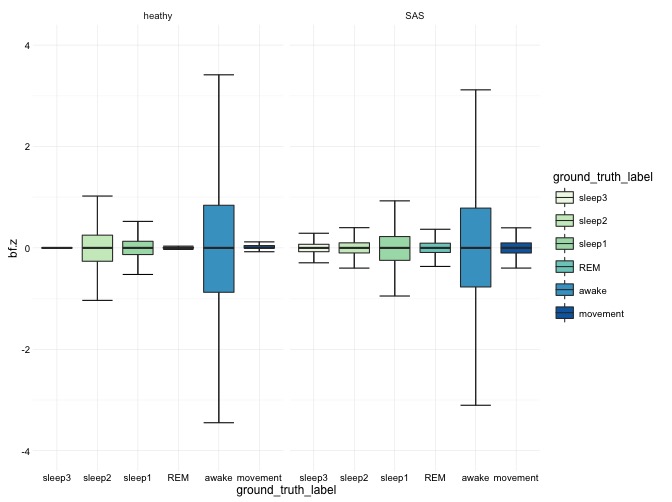


**Figure 1. Visualizing Seep Data.** Ground truth sleep phases determined by polysomnography (top), raw accelerometer values (middle), and light sensor (bottom) from a healthy patient.

**2.3 Exploratory Data Analysis**

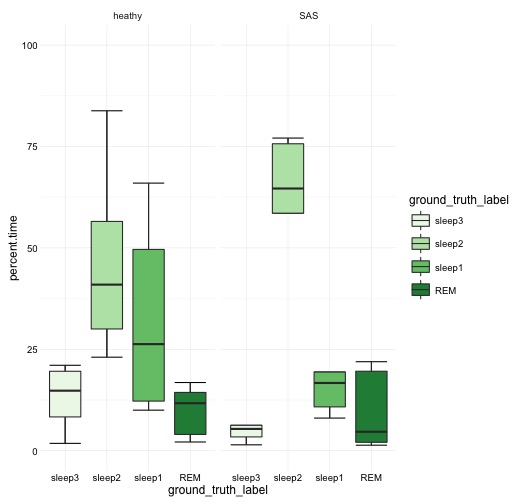
Data from 5 healthy patients and 5 individuals suffering from sleep apnea syndrome (SAS) were analyzed for variation in z-signal and time spent in different sleep phases.

Interestingly, variation of the z-signal (with or without Butterworth filter, described below) showed no differences based on disorder. As expected, episodes of “awake” had the most variation (Figure 2). During REM, a person’s body becomes relatively paralyzed. When looking at movement during REM sleep phases, Figure 2 demonstrates minimal error and a tight interquartile range.



**Figure 2. Boxplots for z-axis movement.** Butterworth filtered z-signal for healthy (n=5) and SAS patients (n=5) based on sleep phase.

Next, we investigated whether SAS affected the amount of time spent in each sleep phase. The total time spent sleeping was obtained for each patient. Then the duration of time spent in each sleep phase was calculated. Based on two-way ANOVA, SAS has no affect on the percentage of time spent in each sleep phase (Figure 3).

****

**Figure 3.** **Percent time spent in sleep phase.** Total time spent in each sleep phase was divided by total sleep time for healthy (n=5) and SAS patients (n=5). Results are presented as percentages.

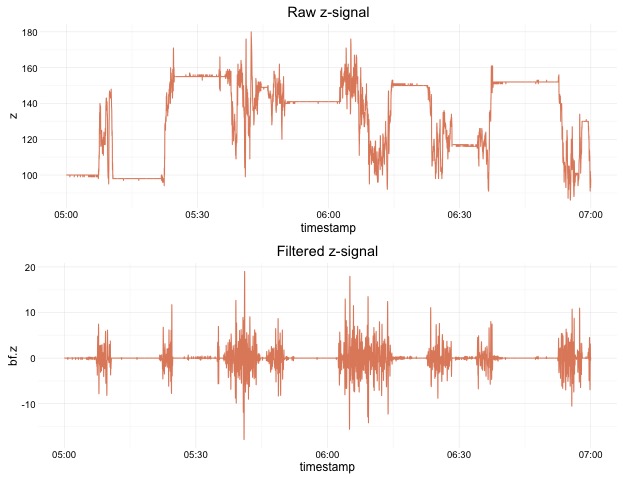
As a means to better characterize differences in sleep quality from different patients, we performed a rigorous extraction of features detailed below.

**3. Feature Engineering**

*Summary:* Data was downloaded, converted to .csv and used for analysis. Upon importing files into R, column names are added and the Unix timestamp is assigned. A series of functions are applied to extract features. First, a Butterworth filter is used to normalize raw 3D measurements. Then metric features, Fourier Transform, and baseline predictions were calculated.

**3.1 Butterworth Filter**

Signal processing filters are designed to reduce or enhance desired aspects of a signal. We applied a 1st order Butterworth IIR filter with bandwidth 3-11 Hz to remove noise present in the data set and to normalize raw x, y, and z measurements. Normalized features are *bf.x, bf.y,* and *bf.z*, respectively. The 3-11 Hz low- and high- cutoff frequencies have been experimentally validated for sleep/wake phase detection using this filter. Filtered data points were used herein.

****

**Figure 4. Results from filtering raw z-signal.** A Butterworth IRR filter was applied as described above to normalize the z-axis.

**3.2 Metric Features**

***Time interval:*** Assessing metrics over distinct epochs, or time periods, in sleep data sets is routine. However, there is no consensus as to which time interval is most accurate for differentiating between stationary and active movements. Published studies employ a vast range of intervals as frequent as every second and as long as 10-minute windows. We selected 15, 30, and 90-second intervals due to the common prevalence of these time periods for activity recognition in literature.

For each time interval, the following metrics were calculated:

* *run.sd –* Running Standard Deviation (always double the metric feature window)
* *int\_mean –* Interval mean
* *int\_max –* Interval maximum
* *int\_min –* Interval minimum
* *int\_sd –* Standard deviation within the interval
* *range –* Interval range (max – min)
* *ratio –* Interval ratio (max / min)
* *int\_RMS –* Root mean squared of the interval
* *bfz.absolute –* Absolute value vector of the normalized uniaxial *z* axis data (bf.z)

**3.3 Fourier Transform**

Fast Fourier transform (FFT) is commonly applied to simulation data as a means to identify the presence of cyclical behavior. FFT was applied to identify the dominant frequency (*freq*) and respective magnitude (*magnitude*) within each time segment.

**3.4 Baseline Activity Predictions**

Simple baseline model predictions (*base.pred*) for active versus sedentary behavior were calculated by the following logic:

where is the running standard deviation with a 30 second window (*run\_sd*). Baseline predictions were around 55% accurate. Using baseline predictions, the duration of activity cycles were calculated in seconds (*duration*) and normalized (sqrt(*duration*) = *duration.norm*. Interestingly, despite it’s unimpressive predictive power the duration of active periods based on the simple predictions are the most useful feature for model accuracy using Random Forest models (detailed below).

**4. Predictive Modeling**

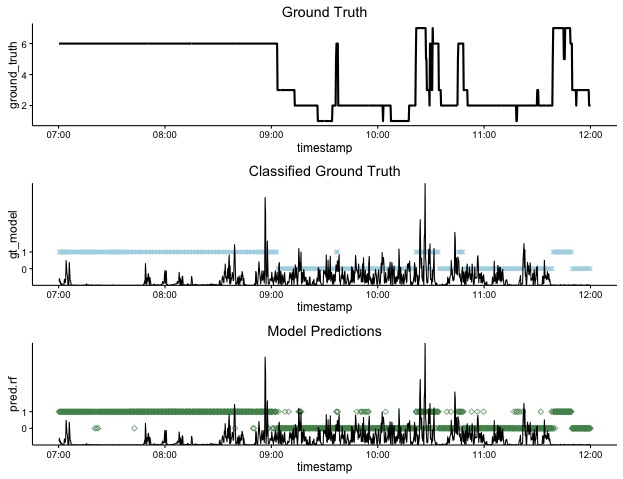
**4.1 Pre-processing data**

For training purposes, time segments with “unknown” activity labels (ground truth = 0) were removed. Furthermore, ground truth labels were simplified to binary classifiers representing sleep (1, 2, 3, or 5) or awake (6 and 7).

**4.2 Random forest models with cross validation**

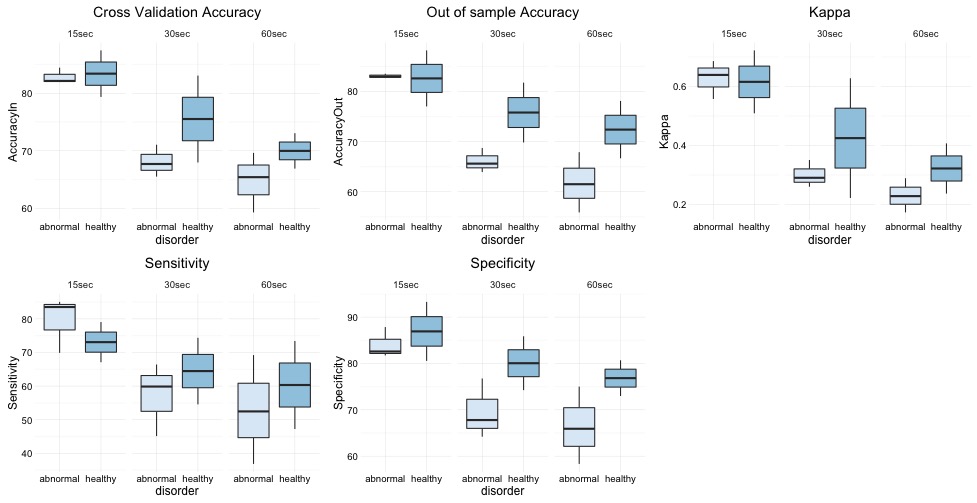
*Data* – Random forest models were selected for classification modeling because they out performed logistic regression models o the same data (data not shown). All random forest models were built using 10-fold cross validation. Data was split 70/30 for training/testing. After a series of iterations, the following features produced the best models (as measured by accuracy) and were used herein: *bfz.absolute*, *duration.norm*, and *magnitude*.

*One-for-all model –* Predictive models demonstrated that training data performed best when personalized (e.g. build one model per person). When training the model with a data set combining healthy and SAS individuals (5 each) using metric features calculated for a 15-second windows, out of sample performance averaged around 70% on test split data(data not shown). Furthermore, introducing completely new patient data (e.g. none of the patient’s data was used for training), the model performed with about 50% accuracy. However, a model created per patient significantly improved accuracy to >80% for nearly all individuals regardless of health status. This suggests that for sleep estimation, a personalized approach must be used.



**Figure 5. Predicting sleep/wake cycles using accelerometer data.** Ground truth PSG readings (top). Ground truth classified into sleep (0) or awake (1) in blue dots plotted over absolute value of filtered z (middle). Predictions from random forest model (green dots, 0 = sleep, 1 = awake) plotted over absolute filtered z (bottom).

*Interval size models* – Predictive modeling also demonstrated a clear advantage in smaller time intervals for feature engineering. Training models for individuals using 30-second and 60-second intervals resulted in model accuracy around 70% and 50% respectively for both in and out of sample predictions on personalized data (data not shown). By reducing the sampling interval to 15 seconds we improved out of sample accuracy to >80% for nearly every individual. To note, individualized models poorly predicted sleep phases for other patients, regardless of health status. Interestingly, predictions for SAS patients suffered more severely when interval size windows were altered, especially with respect to accuracy and specificity. Lack of specificity in this case further demonstrates that sleep quality is underestimated, especially in patients with a sleep disorder.



**Figure 6. Shorter time intervals result in better model performance.** A best random forest model was obtained on training data using a 15sec, 30sec, or 60sec window via 10-fold cross-validation for healthy (n=2) and SAS patients (n=3). Boxplots represent the average performance for the best model for each individual for each time, grouped by disease state.

**5. Conclusions and Future Direction**

With the widespread use of fitness devices, better sleep estimation on actigraphy data is beneficial for everyone from hobby trackers to clinicians.

**5.1 Modeling Summary**

Take home lessons from modeling:

* Sleep predictions are most accurate when performed in a personalized manner
* Assessing properties of sleep (model features) more frequently results in higher accuracy
* Sleep quality is underestimated for individuals with sleep disorders

**5.2 Suggestions for improvements**

Options to improve predictive models:

* Test different digital filters
* Test different types of models (support vector machine)
* Assess the behavior of features before and after sampling window
* Add heart rate monitoring to model
* Add body temperature measurements to fitness watches

**5.3 Code**

The code used for analysis and modeling in this project can be found online at:

https://github.com/cwphilipson/capstone