Drew Hill
PH290 – Project 4, Parkinson's & Smartphone Data

Summary of a Team's Entry

I chose to analyze the entry of Boussios et al, titled "The Construction of a Novel Statistical Algorithm to Objectively Diagnose Parkinson's Disease (PD) Using Smart Phone Data." They created their own PD-diagnostic algorithm, "Forest of Shrubs," to maximize signal from datasets with large data samples from a small number of individuals and take advantage of dependency between longitudinal observations. To inform their algorithm, they assembled and explored a set of 5-minute data chunks representative of a single patient's PD tremor. Using summary statistics gleaned from this dataset, they developed a boosted decision tree based on the amplitude and duration of accelerometer signals. To avoid intra-patient dependency, they randomly divided data into hundreds of 16-row datasets (1 row for each patient) and fit, to each dataset, a small decision tree ("Shrub"). Each shrub was then allowed a vote on the likelihood of disease status, and the average probability of all of the votes cast for each patient was used to determine the final diagnosis. This method was cross validated via 16-fold CV, where each fold represented the rows belonging to a single patient (i.e. CV by dropping one patient's data each time). Ultimately, no model showed predictive power during CV, and thus, the forest of shrubs model was considered inadequate for prediction. A thorough analysis of the distributions of accelerometer data suggested the data were simply too noisy to detect a PD-differentiable pattern in the signals; the authors infer that much of this could be fixed with a more-refined data collection approach with considerations for time of medication, cell phone placement on the body, and baseline testing. Interestingly, they were able to significantly predict "progression" (or time since onset) using linear regression of various summary features of the data.

Summary of the data and my analysis

I explored 1) the challenge question of whether PD can be classified using cellular phone data, specifically accelerometer data and 2) my own question of whether GPS-measured distance travelled can be predicted with accelerometer data. The data as analyzed for the challenge question consisted of 14,580,699 observations of 36 variables (accelerometer, GPS, and personal health data) across 16 patients. From these data, a subset of 80,000 rows of data for 5 variables of interest across 4 male participants was employed for exploration of my own question. Specific reasoning for this subset as well as related data handling details are discussed in later sections of this report.

An initial exploration of the accelerometer data lead to the selection of two features of interest: the frequency of a 15-second averaged value of the "number of samples taken per second" (Nsamp) that were between 15-25 samples/second (15-SAN), and the 5-minute average Signal Magnitude Area (SMA) (Figures 1-3). 15-SAN were highly significantly different between PD and control subjects ($p < 2.2e-16$), with a 15-SAN accounting for, on average, approximately 0.2% of all 15-second average Nsamp values in PD patients and a 15-SAN accounting for approximately 10% of all 15-second Nsamp values for controls (Figure 4). This, in itself, is a very interesting finding, and suggests accelerometer data may be apt at detecting PD patients. Logistic regression of 15-SAN and SMA run on 15-second chunks of data (while controlling for sex) also proved highly useful for classifying PD (Figure 4). For example, controlling for gender, the log odds of a patient having PD increases by -0.30 for every full point increase in the SMA5 of their 15-second data chunk, and -6.11 for if that chunk is fully within the 15-SAN range. The classification power of this set of features on PD status was further demonstrated by the results of a cross-validated, deep-learning algorithm that I performed, which produced an AUROC of 0.82 when applied to 10,000 randomly sampled chunks of data from patients excluded from the learning process (Figure 5). Overall, these analyses demonstrated that accelerometer data can be used to predict PD status with a good deal of accuracy.

A linear regression trained on 3 of the 4 participants that were considered demonstrated a marginally significant relationship between both SMA5 & distance and Number per Sample & distance, in males when controlling for PD (Figure 6b), but performed very poorly in the prediction of distance in the participant that was excluded from the model training set (Figure 6c). I was thereby unable to demonstrate a significant predictive relationship for distance using the accelerometer features of SMA5 and Number per Sample. However, I have hope that a more sophisticated statistical analysis could prove fruitful, especially if able to account for the amount of time spent in a vehicle vs. on foot—this was a major limitation of my analysis.

Data Management

I chose to examine only accelerometer and GPS location data. In order to avoid major memory and hard disk issues, I wrote a .bash script to look into the tarball of raw files, loop through each file, and extract only files with names matching the patterns "*accel*" or "*gps*". Data for one patient of each disease status (Parkinson's and "control") were read into R in order to understand, using a manageably sized subset, what the data generally looked like in terms of variables and data completeness. Data seemed complete, and, aside from id and time stamp, I decided to select the variables for the x, y, and z mean accelerometer data; x, y, and z "PSD.1" data, which corresponds to low-frequency Power Spectral Density data, which the literature suggests correspond well to human movement (Bayat, Pomplun, and Tran 2014); number of accelerometer samples per second; latitude and longitude; and altitude.

Drew Hill

PH290 – Project 4, Parkinson’s & Smartphone Data

Challenge Question: Associate Parkinson’s Disease with Accelerometer Data

A review of the literature indicated that there are no truly ideal ways to summarize accelerometer data, especially in light of the fact that participants in this study were not instructed to carry their phone with them in standardized fashion (i.e. it’s difficult to know when the phone was right-side-up, up-side down, etc.). One promising means of extracting useful information from accelerometer data was found in (Witowski et al. 2014), which examined the average number of accelerometer samples for each continuous 15 second chunk of data (“15-SAN”). In applying this method to the Parkinson’s data, I discovered a unique spike in frequency of 15-SAN between 75 and 85 (Figure 1). Plotting this value by participant demonstrated that this was just an artifact of a high frequency due to a single participant (Figure 2), but it also showed that a spike in density between 15-25 that looked shared, might actually be unique to control subjects. As this makes sense (less movement in controls), I decided to isolate it as a feature (total number of 15-SAN between 15-25).

$$\text{SMA} = \sum_{i=1}^{N}(|x(i)|) + (|y(i)|) + (|z(i)|)$$

Equation 1.

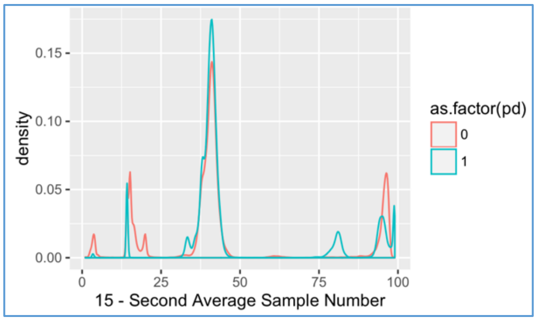


Figure 1. Kernel Density of Sample Number per Second by Disease Status

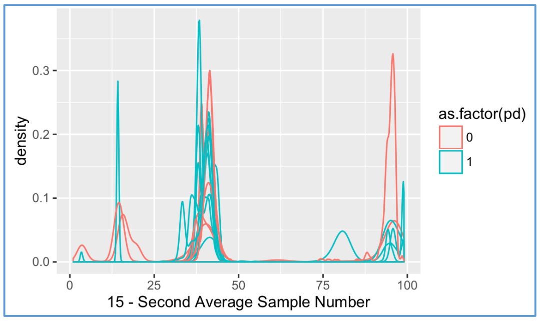
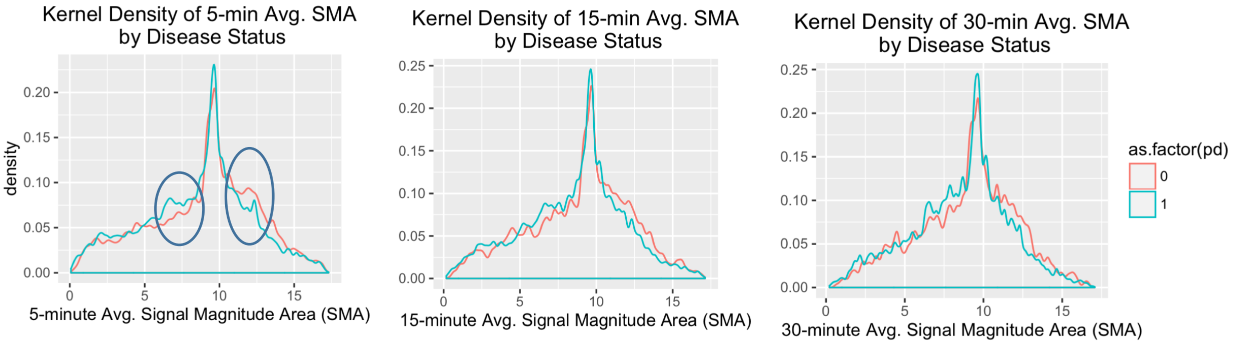


Figure 2. Kernel Density of Sample Number per Second by Disease Status and Participant

I briefly examined x, y, and z coordinates on their own, but the data were messy. I consolidated these data into Signal Magnitude Area (SMA), described in (Khan et al. 2010) and shown in Equation 1, where x(i), y(i), and z(i) indicate acceleration signal along the x, y, and z axes, respectively. SMA magnitude is likely to correspond to different intensities of exercise, and so I began averaging by small chunks of team (about as much as standard exercise might take) to see if anything showed up. looked at 5-minute, 15-minute, and 30-minute average SMA (Figure 3), which showed a slightly different pattern between disease status—higher frequency before the major peak in PD patients, and higher frequency after the major peak in PD patients. Based on this, I decided to include 5-minute average SMA as a feature.

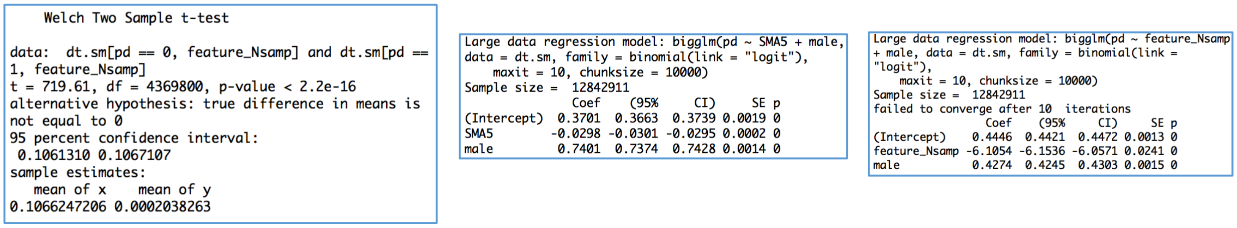

Kernel Density of 5-min Avg. SMA by Disease Status

Kernel Density of 15-min Avg. SMA by Disease Status

Kernel Density of 30-min Avg. SMA by Disease Status

Figure 3. Kernal Density of SMA averaged over different time horizons.

A simple unpaired t-test of the average SAN-15 among controls vs. PD patients demonstrated a very strong effect (p < 2.2e-16), with the mean value for PD patients 0.0002 and the mean value of control patients at 0.1066 (95%CI: 0.1061 – 0.1067) (Figure 4). A logistic regression using “bigglm” with “logit” link showed strong correlation between both SMA & PD when controlling for gender (p<0.00001), and 15-SAN & PD when controlling for gender (p<0.00001). The results are summarized above.



```
	Welch Two Sample t-test

data:  dt.sm[pd == 0, feature_Nsamp] and dt.sm[pd ==
1, feature_Nsamp]
t = 719.61, df = 4369800, p-value < 2.2e-16
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 0.1061310 0.1067107
sample estimates:
   mean of x    mean of y
0.1066247206 0.0002038263
```

```
Large data regression model: bigglm(pd ~ SMA5 + male,
data = dt.sm, family = binomial(link = "logit"),
    maxit = 10, chunksize = 10000)
Sample size =  12842911
               Coef    (95%     CI)     SE p
(Intercept)  0.3701  0.3663  0.3739 0.0019 0
SMA5        -0.0298 -0.0301 -0.0295 0.0002 0
male         0.7401  0.7374  0.7428 0.0014 0
```

```
Large data regression model: bigglm(pd ~ feature_Nsamp
+ male, data = dt.sm, family = binomial(link =
"logit"),
    maxit = 10, chunksize = 10000)
Sample size =  12842911
failed to converge after 10  iterations
                Coef    (95%     CI)     SE p
(Intercept)    0.4446  0.4421  0.4472 0.0013 0
feature_Nsamp -6.1054 -6.1536 -6.0571 0.0241 0
male           0.4274  0.4245  0.4303 0.0015 0
```

Figure 4. T-test for mean 15-SAN by PD status; logistic regression of SMA-5 (middle) and 15-SAN (left), controlling for sex.

To further understand the link between PD and these features and to improve my skills using a package with which I am hoping to become increasingly facile, I decided to take these features to "H2O", the deep learning machine package that I described during my in-class presentation. For this analysis, I setup a deep neural network to run 50 full epochs, 10000 data points at a time with 200-point random drop-out and a hyperbolic tangent (tanH) activation function. The H2O command included an option for cross-validation using the data from 3 patients (Apple, Cherry, and Crocus), two of which were PD patients.
I tested the prediction capability of this model on 10,000 randomly selected 15-second data chunks from the 3 omitted participants (this is the chunk size upon which the model was trained). The model did surprisingly well with an AUROC of 0.8242 (Figure 5).

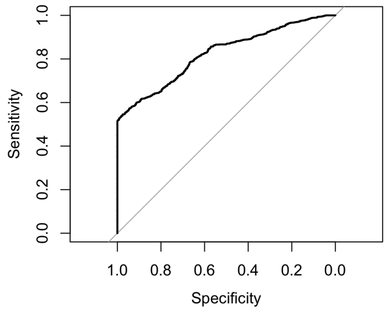
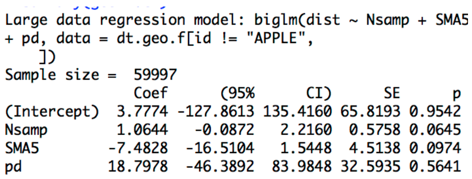


Figure 5. ROC for classification of PD using Deep Learning Model on 10,000 random chunks.

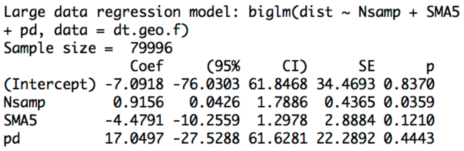
My Question: Can GPS distance be predicted with accelerometer data?

For memory considerations, analysis was performed on only the first 20,000 rows (NA's not included) of 4 participants' data sets (2 PD and 2 control, all males). The analysis was performed as follows: 1) A function obtained from Stack Exchange (and scrutinized by me) was used to produce from the latitude and longitude values a set of rows corresponding to the "n+1"th latitude and longitude values. 2) From these values, the "geospace" package was employed to produce a distance matrix-- using the Haversine function for computational efficiency. This was by far the most ram-intensive step, resulting in ~ 13GB of RAM use per iteration. 3) From this matrix, a vector of the distance (in meters) travelled between each interval was calculated. 4) This new vector was reapplied to the original 20,000-row data set. 5) The 20,000-row datasets for each participant were combined into a single data table for analysis. This ultimately resulted in 79,996 complete row for 5 variables of interest (id, Number per Sample [the precursor to 15-SNA], SMA5, sex, and distance) across 2 PD participants and 2 control participants. This dataset was fed into biglm, and produced a significant relationship between only Number per Sample (Nsamp) and distance travelled, while controlling for PD and SMA5 (p=0.4; Figure 6a). biglm was again used to produce a cross-validated linear equation for distance on only 3 participants (Figure 6b). This model resulted in no significant relationships with distance; Nsamp and SMA5 became marginally significant (p=0.06 and 0.10, respectively). This model suggested that the distance travelled by a male wearing an accelerometer would increase by 0.92 meters with every increase in Nsamp when controlling for PD, decrease by 4.5 meters with every full point increase in SMA5 when controlling for PD, and would be, on average, 18.8 meters greater for those with PD than those without when controlling for SMA5 and Nsamp, however the coefficient for PD, which makes very little sense, was highly insignificant. Notably, when the latter model was used to predict distance in the omitted participant, it performed very badly (Figure 6c).



```
Large data regression model: biglm(dist ~ Nsamp + SMA5
+ pd, data = dt.geo.f[id != "APPLE",
    ])
Sample size =  59997
               Coef      (95%      CI)      SE      p
(Intercept)  3.7774 -127.8613 135.4160 65.8193 0.9542
Nsamp        1.0644   -0.0872   2.2160  0.5758 0.0645
SMA5        -7.4828  -16.5104   1.5448  4.5138 0.0974
pd          18.7978  -46.3892  83.9848 32.5935 0.5641
```

Figure 6a. Linear model predicting distance from accelerometers, all 4 participants.



```
Large data regression model: biglm(dist ~ Nsamp + SMA5
+ pd, data = dt.geo.f)
Sample size =  79996
               Coef     (95%     CI)      SE      p
(Intercept) -7.0918 -76.0303 61.8468 34.4693 0.8370
Nsamp        0.9156   0.0426  1.7886  0.4365 0.0359
SMA5        -4.4791 -10.2559  1.2978  2.8884 0.1210
pd          17.0497 -27.5288 61.6281 22.2892 0.4443
```

Figure 6b. Linear model predicting distance from accelerometers, 3/4 participants.

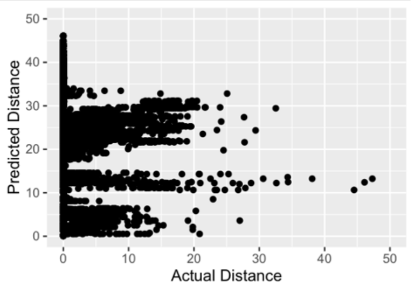


Figure 6c. Plot of predicted distance versus measured distance in the excluded participant, using the model shown in Figure 6b

References

Bayat, Akram, Marc Pomplun, and Duc A. Tran. 2014. "A Study on Human Activity Recognition Using Accelerometer Data from Smartphones." *Procedia Computer Science* 34: 450–57. doi:10.1016/j.procs.2014.07.009.

Khan, Adil Mehmood, Young-Koo Lee, Sungyoung Y Lee, and Tae-Seong Kim. 2010. "A Triaxial Accelerometer-Based Physical-Activity Recognition via Augmented-Signal Features and a Hierarchical Recognizer." *IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE* 14 (5). doi:10.1109/TITB.2010.2051955.

Witowski, Vitali, Ronja Foraita, Yannis Pitsiladis, Iris Pigeot, and Norman Wirsik. 2014. "Using Hidden Markov Models to Improve Quantifying Physical Activity in Accelerometer Data - A Simulation Study." *PLoS ONE* 9 (12): 1–13. doi:10.1371/journal.pone.0114089.