

A Review of the Kaggle Data

And a Proposed Response to the Kaggle Challenge

David Ireland, Simon McBride

19th March 2013 – Version 2.0

Citation

Ireland D, McBride SJ (2013) A Review of the Kaggle Data. CSIRO, Australia.

Copyright and disclaimer

© 2013 CSIRO To the extent permitted by law, all rights are reserved and no part of this publication covered by copyright may be reproduced or copied in any form or by any means except with the written permission of CSIRO.

Important disclaimer

CSIRO advises that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law, CSIRO (including its employees and consultants) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

Contents

1	Introduction	3
1.1	Study Overview	3
1.2	The Competition	4
1.3	Motivation	4
2	Data Analysis	6
2.1	Data Files.....	6
2.2	Non-Sensor Data	6
2.3	Sensor Data.....	7
3	Approaches to a Competition Entry.....	11
3.1	Lifespace	11
3.2	Analysis of Daily Movement and Habits	17
4	Conclusion	24
5	Appendix A	1
6	Glossary	1
7	References.....	2

Figures

Figure 1: Competition data files	6
Figure 2: Medications data mapped to top-level ATC codes	7
Figure 3 Time-series accelerometer data of subject walking with phone in top shirt pocket.	10
Figure 4 PSD as a function of frequency for the time-series in Figure 1.	10
Figure 5 Latitude and longitude for each GPS sample for subjects diagnosed with Parkinson's disease.....	13
Figure 6 Latitude and longitude for each GPS sample for the control subjects.	14
Figure 7 Start and End times for each recording session for control subject Lilly (Age 53)	19
Figure 8 Maximum distance travelled from home-base for control subject Lilly (Age 53).....	19
Figure 9 Start and End times for each recording session for diagnosed PD subject Maple (Age 55; UPDRS 23).....	19
Figure 10 Maximum distance travelled from home-base for PD subject Maple (Age 55; UPDRS 23)	19
Figure 11 Start and End times for each recording session for control subject Sunflower (Age 67).....	20
Figure 12 Maximum distance travelled from home-base for control subject Sunflower (Age 67).....	20
Figure 13 Start and End times for each recording session for diagnosed PD subject Peony (Age 55; UPDRS 14).....	20
Figure 14 Maximum distance travelled from home-base for PD subject Peony (Age 55; UPDRS 14)	20
Figure 15 Start and End times for each recording session for control subject Dafodil (Age 42)	21
Figure 16 Maximum distance travelled from home-base for control subject Dafodil (Age 42)	21
Figure 17 Start and End times for each recording session for diagnosed PD subject Daisy (Age 54; UPDRS 7).....	21
Figure 18 Maximum distance travelled from home-base for PD subject Daisy (Age 54; UPDRS 7).....	21
Figure 19 Heat-map of frequently visited locations for a control subject on a particular day with 10,000m resolution.....	22
Figure 20 Heat-map of frequently visited locations for a control subject on a particular day with 100m resolution.....	22
Figure 21 Heat-map of frequently visited locations for a control subject on a particular day with 100 metre resolution. A enclosed polygon can be used to estimate the area of the subject occupied.....	23

Tables

Table 1: Non-sensor Descriptive Statistics	7
Table 2 Total number of accelerometer and GPS samples at a unique time for subjects diagnosed with Parkinson's disease.....	8
Table 3 Total number of accelerometer and GPS samples at a unique time for the control subjects.	8
Table 4 Life-space analysis for subjects with Parkinson's disease.	15
Table 5 Life-space analysis for control subjects.	16

1 Introduction

In early February of 2013, the Michael J Fox Foundation¹ (MJFF) released a large data-set into the public domain via the Kaggle “Predicting Parkinson's Disease Progression with Smartphone Data” competition². CSIRO, through its ICT Centre’s Australian E-Health Research Centre³ (AEHRC), and UQ, through its Centre for Clinical Research’s Asia-Pacific Centre for Neuromodulation⁴ (APCN), were in the process of negotiating a term sheet⁵ and commencing a project called DeltaMotus when the competition was announced. The DeltaMotus project will investigate the impact of Deep Brain Stimulation (DBS) on Parkinson's disease (PD) by remotely monitoring patient’s pre/post-DBS for motor activity and psychological wellbeing. Both partners considered the MJFF competition a valuable opportunity and decided that their research collaboration would be strengthened by submitting an entry.

This document represents an initial analysis of the competition’s data set and will be used as an input for a joint competition entry.

1.1 Study Overview

The data-set was the result of a study in which a “roughly” age and gender-matched cohort of 16 participants: 9 PD patients at various stages of the disease and 7 healthy controls not manifesting PD at the moment of recruitment. At the beginning of the 8 week study period (which ran between December 2011 and March 2012), the participants were allocated code names in order to obscure their identity and provided the following information:

- Gender
- Current Age
- Age when diagnosed with PD
- Chronic diseases under treatment
- Medications currently being taken

The participants completed a subset of the Unified Parkinson's Disease Rating Scale (UPDRS) at the beginning and end of the study period, which provided two approximate UDPRS scores.

The participants were also asked to carry a smart-phone on their person (in their pocket, in a pouch they wore around their necks, etc) continuously from when they woke until the battery ran out of charge (generally 4-5 hours). The battery was then recharged and the procedure repeated. Most participants would get through a single charge cycle in a day, however several would often stop, recharge, and complete a second cycle in the course of a day. Participants only carried the phones during their awake hours. An application on the phone collected sensor data from the following sensors:

- Microphone (Audio: L1-norm, L2-norm, L-inf norm, power spectral density across four separate bands, 12 lowest mel-frequency cepstral coefficients)
- GPS (Location: latitude, longitude, altitude)
- Accelerometer (Accelerometry: for each of the 3 axes: mean, absolute central moment, standard deviation, maximum deviation, power spectral density across four separate bands)

¹ <https://www.michaeljfox.org/>

² <https://www.kaggle.com/c/predicting-parkinson-s-disease-progression-with-smartphone-data>

³ <http://www.aehrc.com/>

⁴ <http://www.uqccr.uq.edu.au/apcn/>

⁵ Note: This document assumes the term sheet will be agreed with minor additional changes and that it will resolve all questions of intellectual property and other issues of collaboration are covered by the term sheet.

- Compass (Heading: for each of the 3 axes: mean, absolute central moment, standard deviation, maximum deviation)
- Proximity (Presence of objects close to the phone)
- Light meter (Illuminance and luminous emittance as lux)
- Battery level (Charge remaining as a percentage)

1.2 The Competition

Kaggle is a platform that hosts open data prediction competitions. In this instance, the MJFF has posed several challenges for the entrants:

1. Can the data help distinguish PD patients from control subjects?
2. Can the data help to measure the progression, change and/or variability of symptoms in PD subjects?
3. Can the data be used in other creative ways to inform patient treatment, care and/or quality of life?
4. Do the analyses and proposed uses of the data use innovative approaches and methods?

The submission is due on the March 26th 2013, 11:59 EST and must comprise:

1. An entry describing all data/inputs, use of the data, preliminary findings/evidence to support the use of objective, passively collected data points to address the contest objectives and the potential utility if used in a broader way.
2. A narrative explaining what you did with the data (no more than 5 pages). Additional information (images, charts, etc) can be included in the submission as an appendix.
3. There should be an additional page describing your team and contact info.

Entries must be submitted in electronic format via Kaggle's online submission tool. The judging committee includes the following people:

- Alexandra Carmichael - Co-founder of CureTogether and Director of Quantified Self
- Karl E. Case - Professor of Economics Emeritus at Wellesley College
- Maurizio Facheris - Associate Director, Research Programs at The Michael J. Fox Foundation
- Ken Kubota- Director of Kinetics Foundation
- Daniel Vannoni - Entrepreneur and Managing Director of Gecko Ventures

The winning entrant will receive a US\$10,000 prize with no strings attached. The winner will also be invited to present their findings at an MJFF-sponsored event, with the MJFF covering travel and event costs for up to two people.

1.3 Motivation

The collaborative partners' motivations for submitting an entry are:

1. To strengthen the collaboration by:
 - a. Bringing a team together to meet the highly specific, short term goal of submitting a competitive entry to the competition;
 - b. Jointly authoring a journal publication that builds on the entry to provide a broader, multi-disciplinary academic perspective on the competition and our findings;
 - c. Exercising the internal, bureaucratic systems of both organisations to discover potential problems (e.g. intellectual property management, communications, etc) before recommencing the original goals of the collaboration.
2. To use the competition to promote the collaboration as a leading, academically focussed team in the PD community. The communication strategy around the competition is:
 - a. Submission of the entry document (publically available);

- b. Submission of a manuscript to a (possibly open access) academic journal publication;
- c. Press releases by CSIRO and UQ promoting the entry;
- d. A keynote presentation by Prof. Helen Chenery at AEHRC's 2013 E-Health Research Colloquium and subsequent presentations by other members of the team;
- e. Finally, in the event that our entry wins the competition, the profiles of both partners will be significantly enhanced.

Note: Neither partner is motivated by the size of the prize money. Any prize money, minus the US tax deduction, will fund future collaborative research.

2 Data Analysis

2.1 Data Files

The data were supplied to registered users via the competition page on the Kaggle web site in the formats listed at Figure 1.




File Name	Available Formats
HDL	.zip (42.76 kb)
HumDyn	.R (4.79 kb)
HDL Data Documentation	 .pdf (43.76 kb)
Study Overview	 .pdf (49.40 kb)
Participant Codes and Description	 .pdf (48.41 kb)
UPDRS Part 1 Questionnaire-initialscore	.xls (109.00 kb)
UPDRS_Questionnaire_Blank	.docx (30.71 kb)
binary_sample.tar	.bz2 (2.03 kb)
text_sample.tar	.bz2 (2.53 kb)
mjff_binary_files	.zip (6.35 gb)
mjff_text_files	.zip (10.47 gb)
UPDRS Part 1 Questionnaire 2	.xlsx (15.96 kb)
Participant Description	.xls (29.00 kb)

Figure 1: Competition data files

These files were downloaded to and compressed files extracted onto CSIRO computing systems. Where the data were not processed by hand, a collection of R and bash scripts were then written to process the data. The scripts are under version control in the DeltaMotus project Subversion repository⁶.

2.2 Non-Sensor Data

The non-sensor data consists of the following variables (with a data type):

- Gender (M, F)
- Current Age (integer)
- Age when diagnosed with PD (integer)
- Chronic illnesses under treatment & medications currently being taken (free text)

The entire non-sensor data set is at Section 5. A partial summary of the characteristics of this data is at Table 1, with a discussion of chronic illnesses under treatment & medications currently being taken in following sections.

Characteristic	Controls (n=7)	PD (n=9)
Females (%)	2 (28.6)	2 (22.2)
Current Age (mean years, SD years)	61.1 (13.1)	59.5 (10.1)
Age at diagnosis (mean years, SD years)	N/A	51.9 (8.8)

⁶ <https://svnserv.csiro.au/svn/deltamotus/Kaggle/trunk/>

Initial UDPRS score (mean, SD)	N/A	13.3 (7.4)
Follow-up UDPRS score (mean, SD)	N/A	12.6 (6.9)

Table 1: Non-sensor Descriptive Statistics

2.2.1 CHRONIC ILLNESS DATA

All participants were asked to supply details of any non-PD chronic illnesses they suffered. The responses were sketchy and non-standardised, which made analysis difficult. The data suggest a range of non-PD chronic illnesses across the cohort, although 2 PD participants explicitly state no additional illnesses.

2.2.2 MEDICATIONS DATA

All participants were asked to supply details of their medication use. As with the chronic illness data, the responses were sketchy and non-standardised, which made analysis difficult. PD participants reported far more medication use. Figure 2 shows the medications data mapped to top level ATC codes.

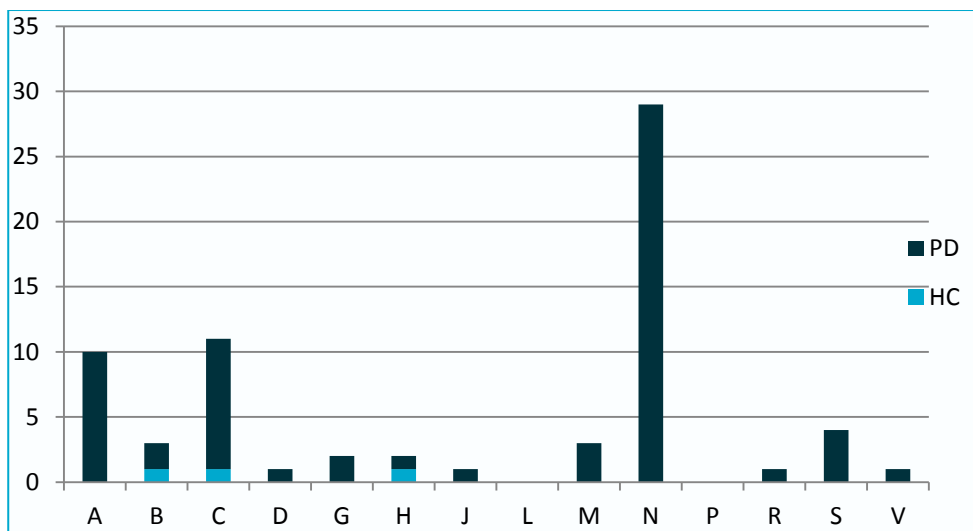


Figure 2: Medications data mapped to top-level ATC codes

Figure 2 suggests PD participants reported more medication use than control participants. The mapping also indicates that medication use by PD participants is skewed towards medications that affect the N (Nervous system), C (Cardiovascular system) and A (Alimentary tract and metabolism) systems.

2.3 Sensor Data

Sensor data consists of:

- Microphone (Audio: L1-norm, L2-norm, L-inf norm, power spectral density across four separate bands, 12 lowest mel-frequency cepstral coefficients)
- GPS (Location: latitude, longitude, altitude)
- Accelerometer (Accelerometry: for each of the 3 axes: mean, absolute central moment, standard deviation, maximum deviation, power spectral density across four separate bands)
- Compass (Heading: for each of the 3 axes: mean, absolute central moment, standard deviation, maximum deviation)
- Proximity (Presence of objects close to the phone)

- Light meter (Illuminance and luminous emittance as lux)
- Battery level (Charge remaining as a percentage)

A review of the source code of the Android application has revealed the application opens a recording session and obtains readings for the sensors periodically for approximately one hour. Subsequently, the obtained data is post-processed and exported to ASCII formatted text files.

Of the given data, the most commonly used towards quantifying PD progression are the recordings from the accelerometer, GPS and microphone (audio) sensors. Each obtained sensor reading was time and date stamped. Table 1 provides the number of obtained samples obtained at unique time and date stamps for each participant over the two month period. These tables show the subjects with Parkinson's disease recorded approximately two as much data as the control subjects. A summary and discussion on each sensor data-set is given in the preceding sections.

Table 2 Total number of accelerometer and GPS samples at a unique time for subjects diagnosed with Parkinson's disease.

SUBJECT	NUMBER OF SAMPLES (ACCELEROMETER)	NUMBER OF SAMPLES (GPS)	NUMBER OF SAMPLES (AUDIO)
Violet	1,189,883	1,304,170	1,450,985
Daisy	1,765,132	578,471	1,765,360
Cherry	317,268	208,438	327,693
Crocus	301,354	211,855	306,226
Orchid	1,342,478	1,363,194	1,878,138
Peony	660,591	742,672	917,796
Iris	73,523	37,709	1,330,943
Flox	938,036	525,255	1,255,330
Maple	1,692,834	661,418	1,824,850
Total	5,281,099	5,633,182	11,057,321

Table 3 Total number of accelerometer and GPS samples at a unique time for the control subjects.

SUBJECT	NUMBER OF SAMPLES (ACCELEROMETER)	NUMBER OF SAMPLES (GPS)	NUMBER OF SAMPLES (AUDIO)
Apple	357,451	179,080	419,651
Dafodil	1,094,890	710,939	1,114,294
Lilly	1,167,307	854,088	1,270,330
Orange	587,972	355,037	608,508
Rose	222,929	20,745	224,628
Sunflower	614,413	223,830	791,752
Sweetpea	210,824	142,663	226,511
Total	4,255,786	2,486,382	4,655,674

2.3.1 AUDIO

The source code of the Android application the participants had the option of allowing the keep the raw audio data recorded or discarding it. Unfortunately, it seems only two participants, Crocus and Sweetpea, allowed the raw audio data to be saved to file. The audio data was encoded in 16-bit, little-endian format and sampled at 8000 Hz. Listening to the available raw audio data can be done using the freely available Audacity software. The post-processed data however is available for each participant. This includes the following:

- Absolute energy of the audio captured
- Root mean square (RMS) energy of the audio captured
- Peak energy of audio captured
- The average, power spectrum density between:
 - 0 - 250 Hz,
 - 250 - 500 Hz,
 - 500 - 1000 Hz
 - 1000 - 2000 Hz.
- 0 – 12 Mel-frequency cepstral coefficients

Although the microphone sensor was turned on periodically, there is no clear indication if the voice of the subject was recorded and furthermore, it is likely significant interference was also recorded i.e. ambient noises.

2.3.2 ACCELEROMETER

A sample from the accelerometer data comprises three values corresponding to the x- y- and z-axis components. The Android operating system typically allows a maximum sampling rate of 50 Hz for the accelerometer sensor. It is presumed the software developers anticipated the limited storage space available and in the Android application followed the procedures:

1. Obtain 50 x-, y- and z-axis accelerometers values approximately every 1 second.
2. Compute the mean, absolute deviation, standard deviation for each axis and record to file.
3. Compute the average power spectrum density between for 0 – 1 Hz, 1 – 3 Hz, 3 – 6 Hz and 6 – 10 Hz and record to file.
4. Compute the average of the 50 time-series samples and record to file.

Step 4 in effect, down-samples the data from 50 Hz to 1 Hz sample rate. A significant amount of information is lost at this step. According to the Nyquist–Shannon sampling theorem a sample rate of 1 Hz has a maximum analysable frequency of 0.5 Hz. To appreciate the reduction of information that is occurring, Figure 1 shows recorded time-series data from an accelerometer sensor embedded in a Motorola X910 running the Android operating system. The phone was located in the top shirt pocket of the subject while they walked casually for approximately 14 seconds. Figure 2 shows the power spectrum density (PSD) as a function of frequency for the time-series data of Figure 1. There are a large number of dominant frequency components greater than 0.5 Hz that would be removed during step 4.

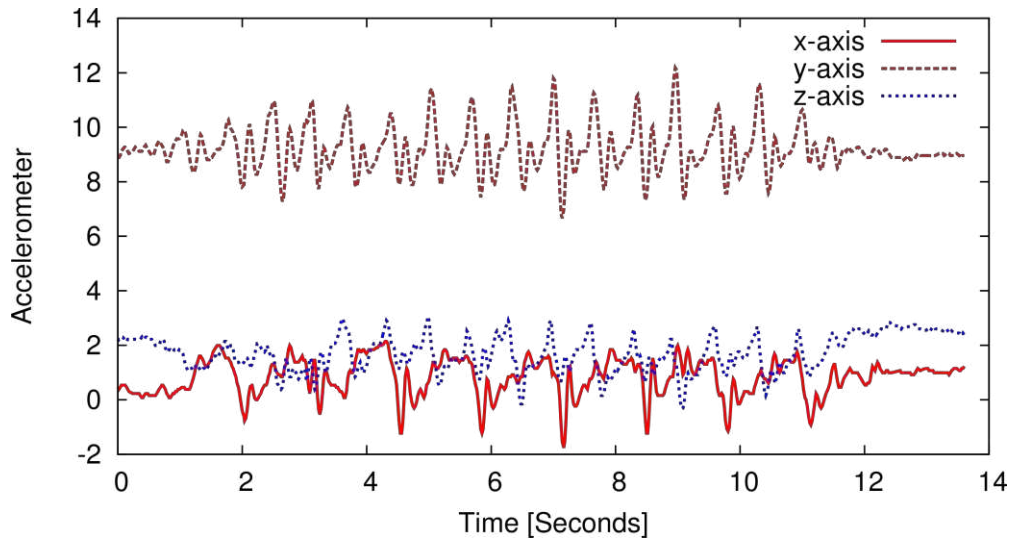


Figure 3 Time-series accelerometer data of subject walking with phone in top shirt pocket.

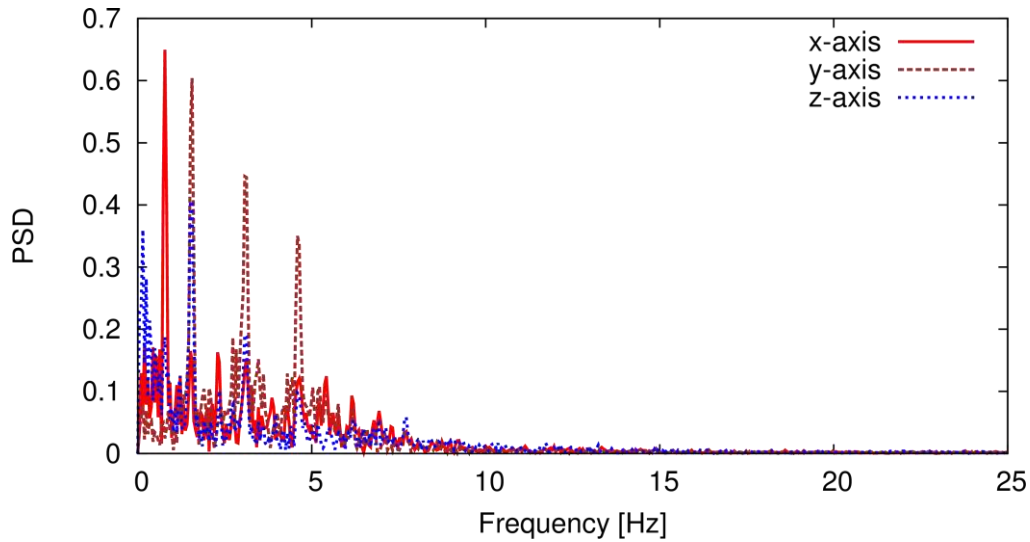


Figure 4 PSD as a function of frequency for the time-series in Figure 1.

2.3.3 GPS

A sample from the GPS sensor comprises two values corresponding to the latitude and longitude of the estimated location. The maximum sample rate of the GPS sensor is dependent on the hardware of the phone and the version of the Android operating system. On average the GPS recordings of this study were sampled at 1 Hz.

The accuracy of the reported latitude and longitude values can vary depending on the number of GPS satellites visible by the sensor, whether the sensor is located indoors or outdoors, ambiguities due to atmospheric changes and interfering signals. Unfortunately, determining the accuracy of the recorded coordinates in this study is arduous. Although the Android operating system has the capability of determining the accuracy of the GPS reading, this was not recorded. However, assuming the GPS sensor was located outdoors, it was quite likely to obtain an average accuracy of approximately 10 metres based on testing done by the authors.

3 Approaches to a Competition Entry

The Data Analysis section provides descriptions of the limitations of the data set. These limitations restrict the range of approaches available when considering an entry to the competition. This section provides an overview of techniques that we believe form a viable basis for an entry.

3.1 Lifespace

An individual's life-space is commonly defined as the size of the spatial area that is travelled through in daily life, as well as the frequency of travel within a specific time frame (Boissy *et al.* 2011). In recent times, quantifying an individual's life-space has been proposed as a more accurate way to capture the functional and psychological aspects of mobility and offering a more accurate metric on actual mobility performance (Boissy *et al.* 2011).

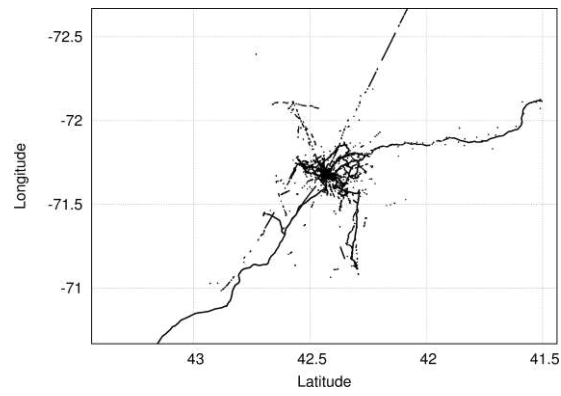
Traditionally life-space assessments were done with subjective, self reports; however, with the advent of ubiquitous smart phones, this process can now be autonomous and likely capture more accurately the actual spatial extent of movement. Furthermore, life-space analysis can be done with a variety of commonly available electronic sensors such as accelerometers, GPS and pedometers. The data-set of this study, in particular the GPS coordinates, provides a rich set of movement activities for each participant. As such given the limited usefulness of the accelerometer and audio data, and the closeness of the submission deadline, it is believed the GPS data is the most useful for mounting a response to the Kaggle challenge. We assert a so-called life-space analysis provides a significant response to the challenges posed.

3.1.1 GENERAL LIFE-SPACE ASSESSMENT

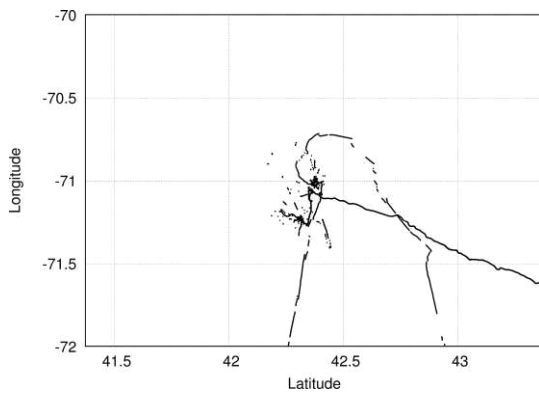
Before providing a life-space assessment it is first required to define a reference point referred as the *home-base*. As this was not specified in the study data it is estimated here as the statistical mode of the recorded GPS latitude and longitude values. Figure 1 gives plots of the latitude and longitude for each participant diagnosed with Parkinson's diseases with their corresponding age and UPDRS score. The home-base location for the particular subject is the centre of each plot. Figure 2 gives plots of the latitude and longitude for each control subject with their corresponding age. Tables 5 and 6 give statistical measures relating to the number of days recorded, **maximum and average distance the subject travelled, and the percentage of time the subject stayed within 100m of their home-base**. This last metric will be referred to as the *home-base score* and is expressed as a percentage.

Referring to Figure 1, there is a **noticeable decrease in recorded activity as the UPDRS of the subject increases**. This is reflected in Tables 5 and 6 showing the subjects with Parkinson's disease commonly stayed in proximity of their home-base 7% more time than the control subjects. Figure 3 gives the home-base score as a function of average UPDRS for each subject. There is evidently a significant linear relationship with a correlation coefficient, denoted r_{xy} , and a coefficient of determination, denoted R^2 , of 0.8074 and 0.65 respectively.

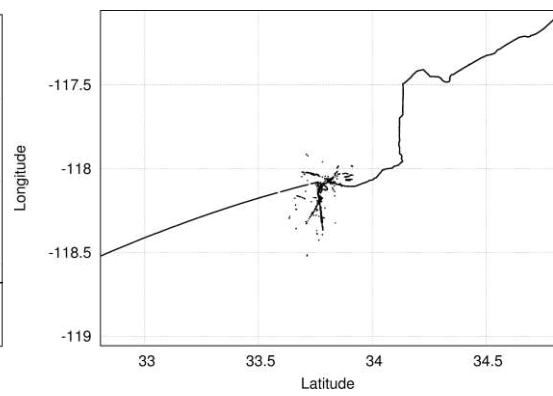
Intuitively one would expect many parameters to affect an individual's desire to stay in proximity of their home-base. For instance age, lifestyle habits and occupation would likely have a significant impact. As age is a parameter of this study Figures 5 and 6 provides the relationship between the home-base score as a function of age for the PD and control subjects respectively. r_{xy} for both of these plots is 0.42 and 0.56 respectively; R^2 is 0.18 and 0.31 respectively. Evidently, there is **more correlation with the age of the control subjects and their home-base score compared to the PD subjects as expected due to the presenting motor symptoms**. In conclusions, despite the small number of PD subjects in this study, the home-base score provides a very simple and potentially useful metric to track disease progression.



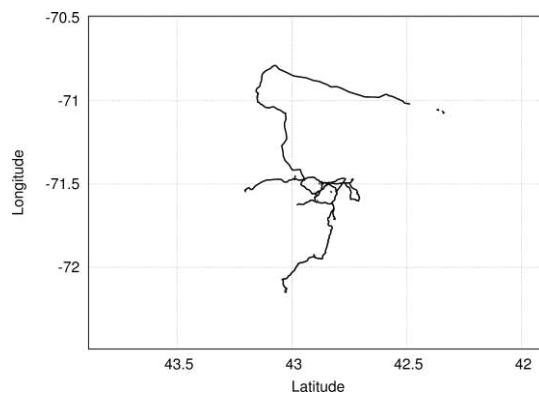
(a) Violet (Age 55; UPDRS 5)



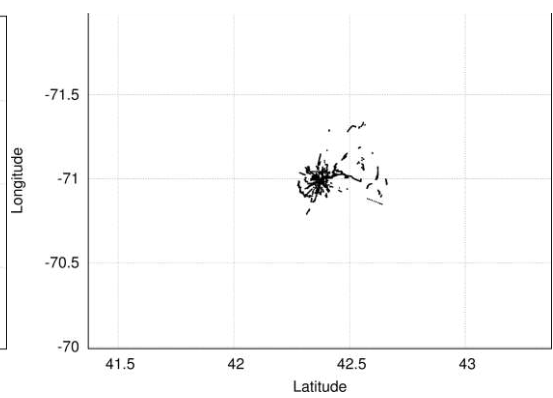
(b) Daisy (Age 54; UPDRS 7)



(b) Cherry (Age 55; UPDRS 7)

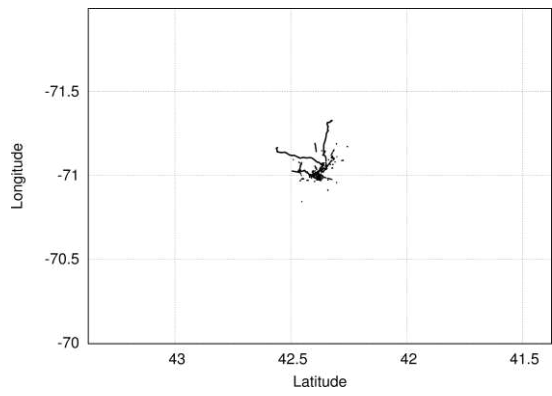


(c) Crocus (Age 46; UPDRS 8)

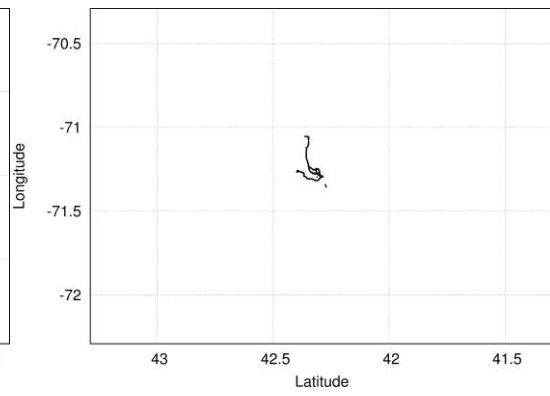


(d) Orchid (Age 69; UPDRS 10)

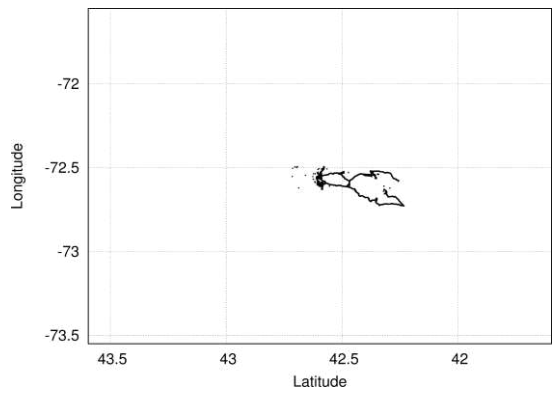
Figure 4 Latitude and longitude for each GPS sample for subjects diagnosed with Parkinson's disease.



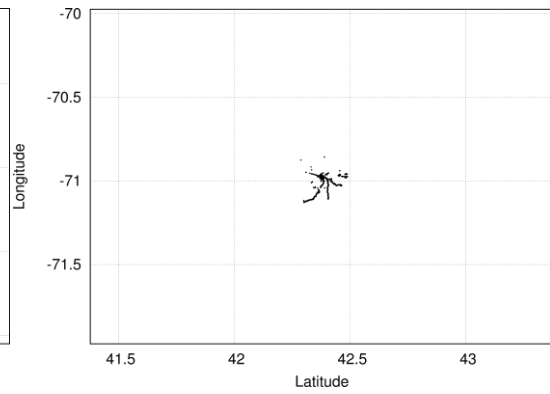
(e) Peony (Age 80; UPDRS 14)



(f) Iris (Age 65; UPDRS 17)

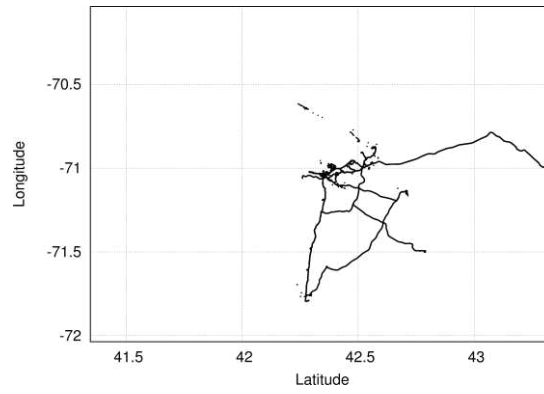


(g) Maple (Age 55; PDRS 23)

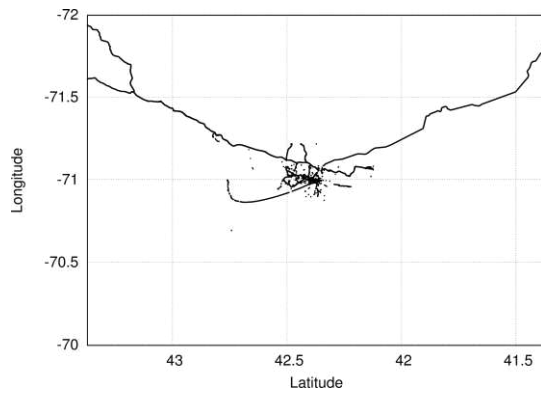


(h) Flox (Age 57; UPDRS 23)

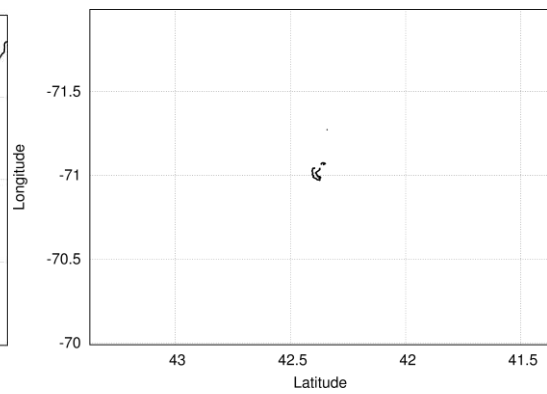
Figure 5 Latitude and longitude for each GPS sample for subjects diagnosed with Parkinson's disease.



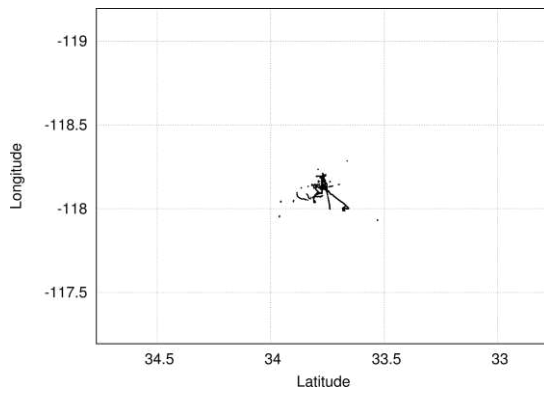
(a) Dafodil (Age 42)



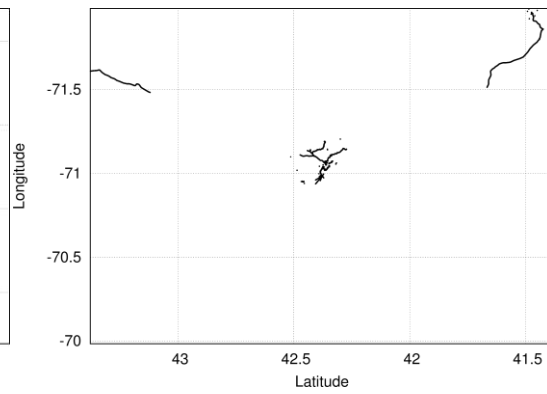
(b) Lilly (Age 53)



(c) Rose (Age 55)

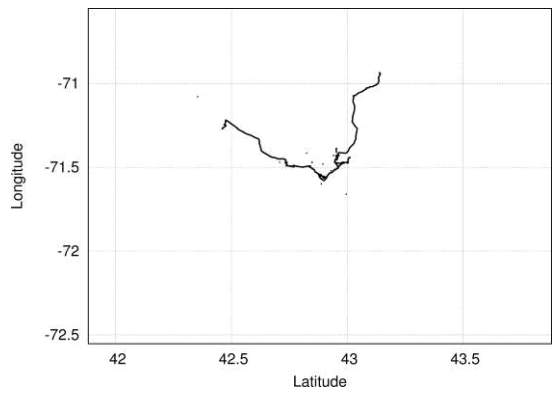


(d) Orange (Age 57)

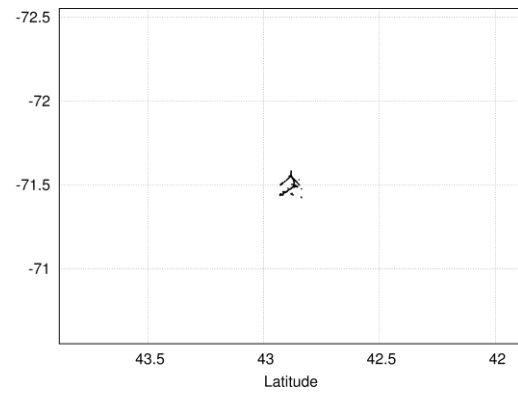


(e) Sunflower (Age 67)

Figure 6 Latitude and longitude for each GPS sample for the control subjects.



(f) Apple (Age 77)



(g) Sweetpea (Age 77)

Figure 5 Latitude and longitude for each GPS sample for control subjects.

Table 4 Life-space analysis for subjects with Parkinson's disease.

SUBJECT	AGE	UPDRS BEFORE STUDY	UPDRS AFTER STUDY	NUMBER OF DAYS RECORDED	MAXIMUM DISTANCE (KM)	MAXIMUM DISTANCE (MILES)	DAILY AVERAGE DISTANCE (KM)	DAILY AVERAGE DISTANCE (MILES)	HOME- BASE SCORE (%)
Violet	55	5	5	64	664	413	34	216	27
Daisy	54	8	7	72	4,161	2585	347	32	20
Cherry	55	7	7	15	316	196	51	10	23
Crocus	46	8	8	20	67	42	16	31	52
Orchid	69	15	10	70	765	475	50	5	56
Peony	80	13	14	43	27	17	8	6	68
Iris	65	16	17	6	21	13	10	6	66
Maple	55	26	23	40	42	26	10	6	62
Flox	57	23	23	28	15	9	4.4	3	79
Mean	60	N/A	N/A	40	675	420	59	37	50

Table 5 Life-space analysis for control subjects.

SUBJECT	AGE	NUMBER OF DAYS RECORDED	MAXIMUM DISTANCE (KM)	MAXMIUM DISTANCE (MILES)	DAILY AVERAGE DISTANCE (KM)	DAILY AVERAGE DISTANCE (MILES)	HOME-BASE SCORE (%)
Apple	77	22	233	145	14	9	58
Dafodil	42	47	215	134	44	27	26
Lilly	53	56	1,140	708	210	130	42
Orange	57	19	35	22	13	8	22
Rose	55	7	23	14	8	5	61
Sunflower	67	37	1,056	656	120	75	34
Sweetpea	77	13	11	7	5	3	56
Mean	61	29	379	235	59	37	43

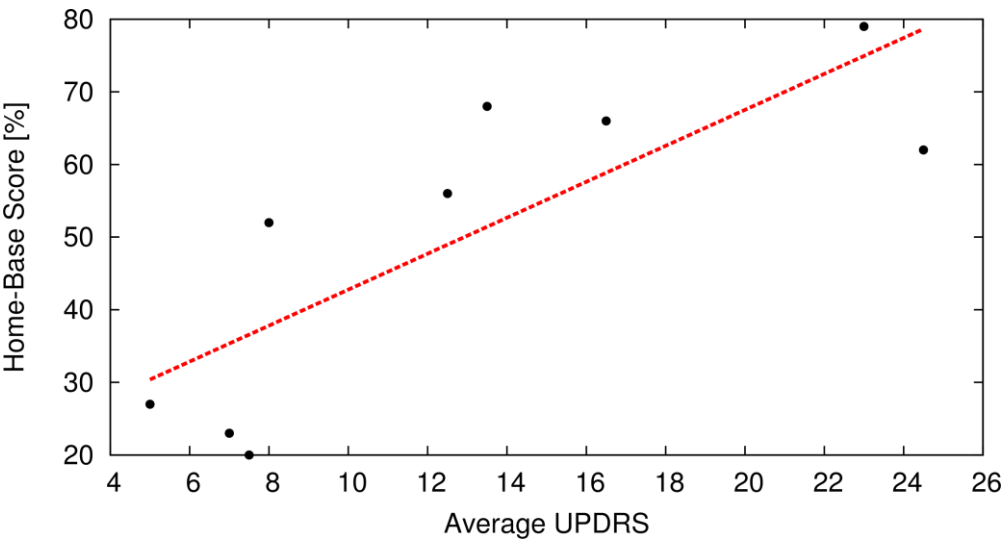


Figure 6 Home-base score as a function of average UPDRS. $r_{xy}=0.80$; $R^2=0.65$

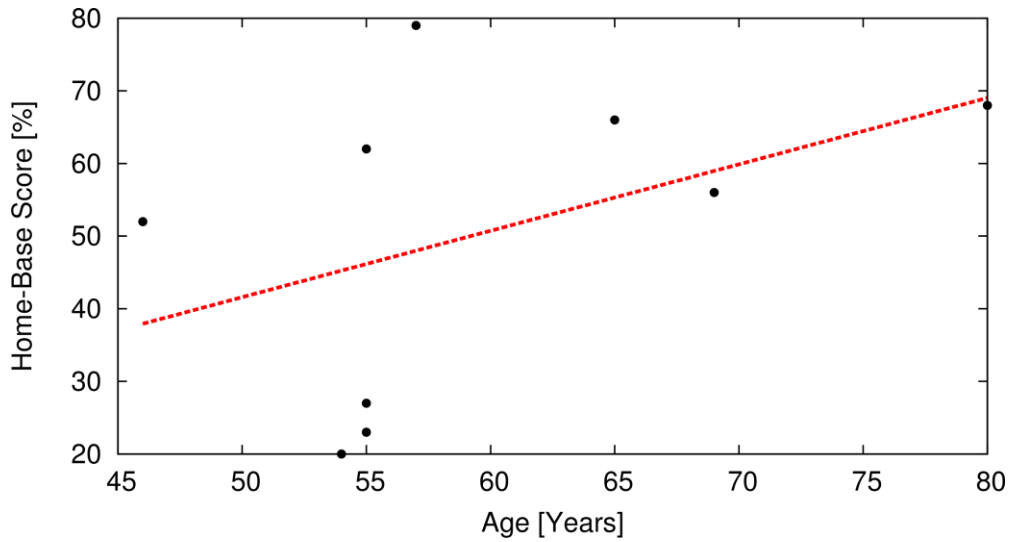


Figure 7 Home-base score as a function of age for the subjects diagnosed with PD. $r_{xy}=0.43$; $R^2=0.18$

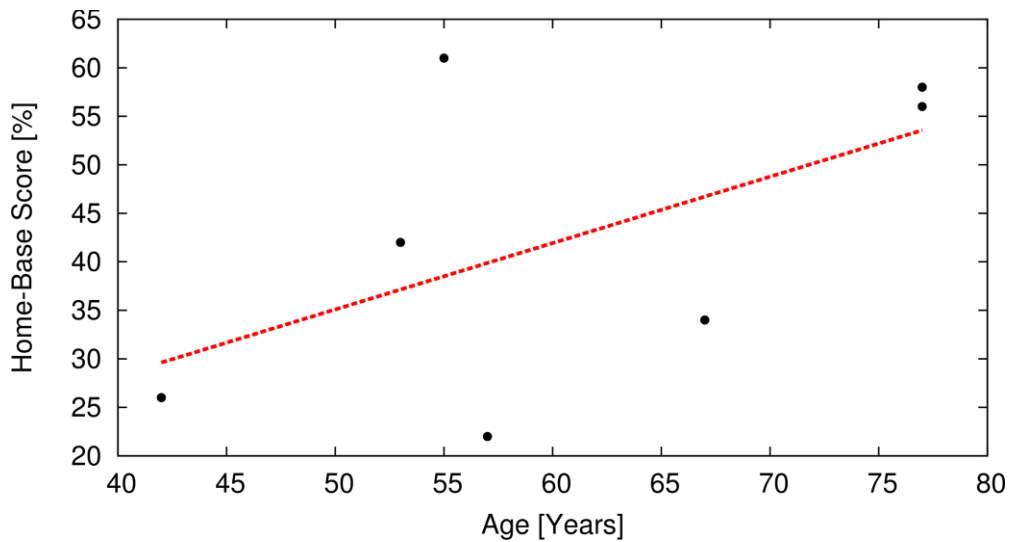


Figure 8 Home-base score as a function of age for the control subjects. $r_{xy}=0.56$; $R^2=0.31$

3.2 Analysis of Daily Movement and Habits

As each sample of data is date and time stamped, analyses of daily activities is possible. Using the GPS data-set, it is also possible to examine the starting and end session times of each day. As mentioned in the study description, the user is required to manually activate the recording application; presumably the application is ended by the user terminating the application or the phone being switched off during the charge cycle. The beginning and ending times of each day might provide some insight into the daily routines of each subject. Figures 7, 9 and 11 provide the beginning and end of the recording sessions for control subjects Lilly, Dafodil and Sunflower respectively. Figures 9, 13 and 17 provide the beginning and end of the recording sessions for PD subjects Maple (UPDRS 23), Peony (UPDRS 14) and Daisy (UPDRS 7) respectively.

These Figures reveal a high periodic routine for subjects Lilly and Dafodil, a moderate periodic routine for Daisy and Peony and a sporadic routine for Maple. Sunflower recorded insufficient data to make any speculative conclusions. Returning back to Maple who scored the highest UPDRS of 23, it is of interest as to

whether the Maple sessions times are a result of Parkinson's disease. In particular, could insomnia be manifesting in this data analysis?

Of further interest is the daily maximum distance the subject travelled from their home-base. Figures 8, 12 and 16 gives the maximum distance in miles for each recorded day for control subjects Lilly, Sunflower and Dafodil respectively. Figures 10, 14 and 18 give the distance PD subjects Maple, Peony and Daisy respectively. There is a notable decrease in maximum distance for PD subjects Peony and Maple presumable due to their higher UPDRS scores, whereas subjects Lilly, Sunflower and Daisy routinely made trips more than 100 miles.

A more fine-grain life-space analysis is also possible that could examine the movements and habits of the subjects comprising: location types, frequency of visits, modes of transport. Unfortunately at present there is no autonomous software tool available and thus an exhaustive human analysis is required.

However, a semi-autonomous software tool has been developed that enables the GPS coordinates of the data to be overlayed on top of a Google map image. This would allow notable places, roads and street views to be also identified. Visualisation in the form of a heat-map is implemented where the higher the intensity (red) the more frequent the location is visited for that particular day. This is expected to be directly correlated to the time the subject stays at that location.

Figure 19 gives an example of a control subject that made a journey from their home-base to a sporting centre. As the route the subject took was via a road, one can deduce the mode of transportation was by car or bus as opposed to a train or tram network.

As the user zooms into the map, the resolution of space travelled through by the subject can be increased. Figure 20 gives a zoomed in image over the sports centre comprising the car park (presumable by the P symbol) and the Lundholm gym. Curiously the subject has evidently stayed longer in the car park than the gym. It is expected a higher resolution Google and heat-map might provide more information, but for brevity will not be shown here. Of course there are a plethora of ethical and privacy concerns with this form of analysis that will have to be discussed prior to the Kaggle submission.

A common analysis in determining the life-space of a person is to determine the surface area the subject travels through rather than a distance measure. Of course this would require modern electronic sensors and therefore has only been recently investigated, see for example (Boissy *et al.* 2011). Given the GPS data-set and the developed heat-map tool it is possible to determine this area. Figure 21 provides an example of this, where the user can draw a polygon shape around the heat-map area allowing the surface area (in square metres) to be estimated.

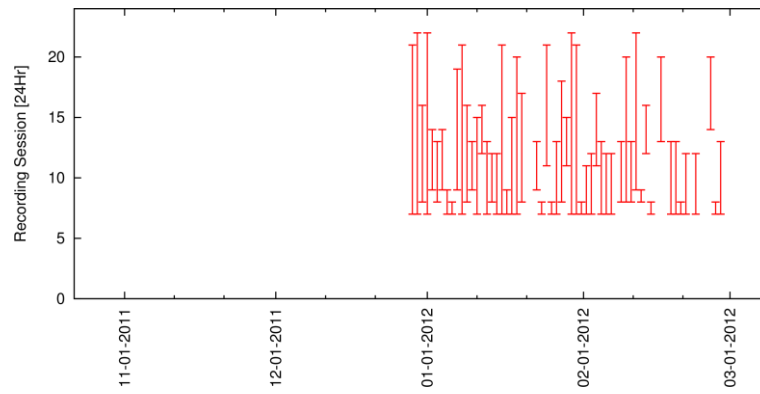


Figure 7 Start and End times for each recording session for control subject Lilly (Age 53)

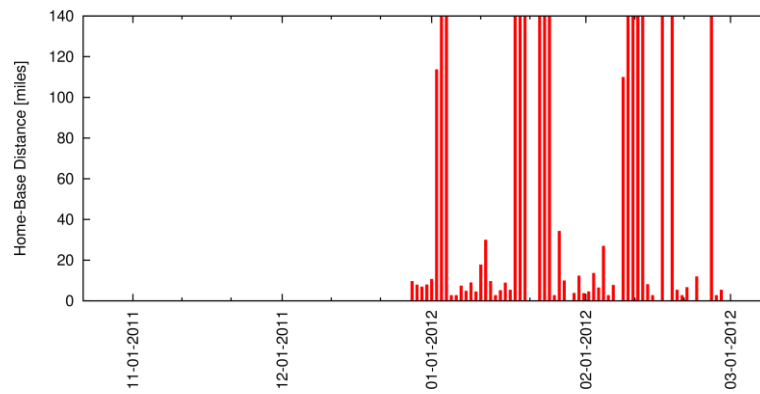


Figure 8 Maximum distance travelled from home-base for control subject Lilly (Age 53)

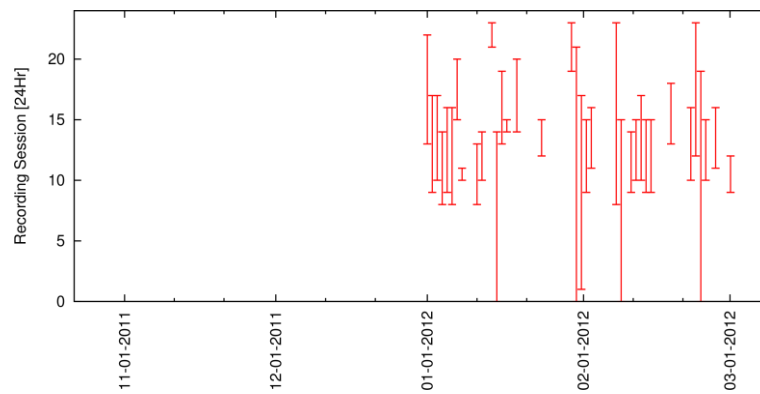


Figure 9 Start and End times for each recording session for diagnosed PD subject Maple (Age 55; UPDRS 23)

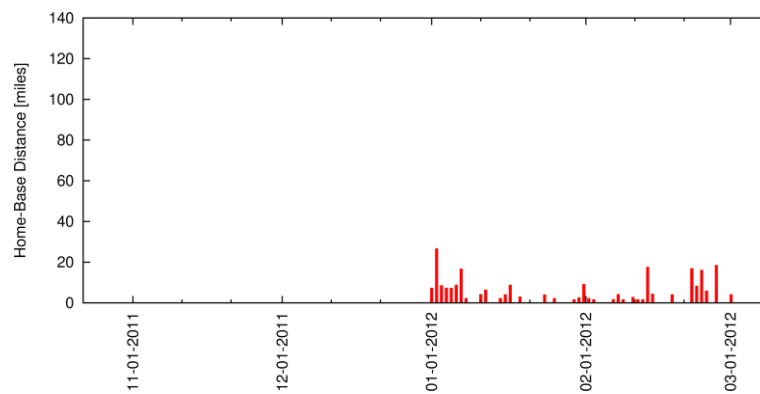


Figure 10 Maximum distance travelled from home-base for PD subject Maple (Age 55; UPDRS 23)

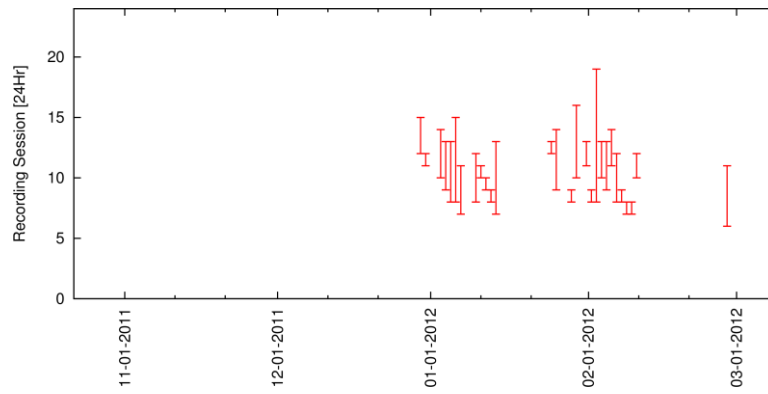


Figure 11 Start and End times for each recording session for control subject Sunflower (Age 67)

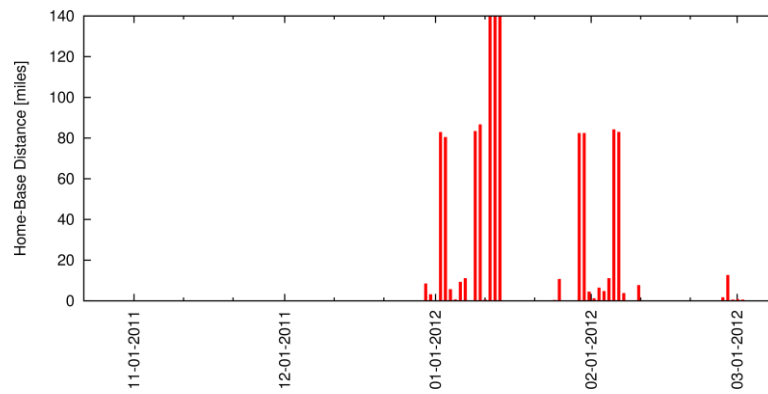


Figure 12 Maximum distance travelled from home-base for control subject Sunflower (Age 67)

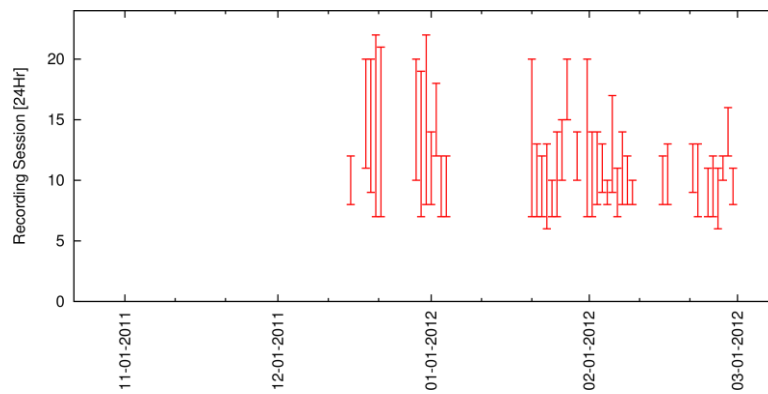


Figure 13 Start and End times for each recording session for diagnosed PD subject Peony (Age 55; UPDRS 14)

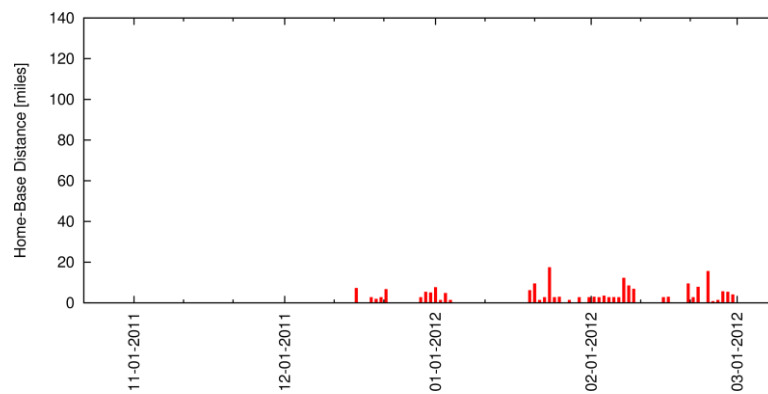


Figure 14 Maximum distance travelled from home-base for PD subject Peony (Age 55; UPDRS 14)

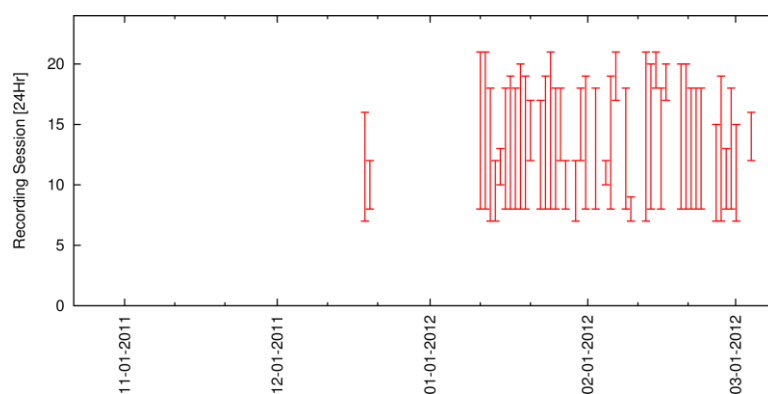


Figure 15 Start and End times for each recording session for control subject Dafodil (Age 42)

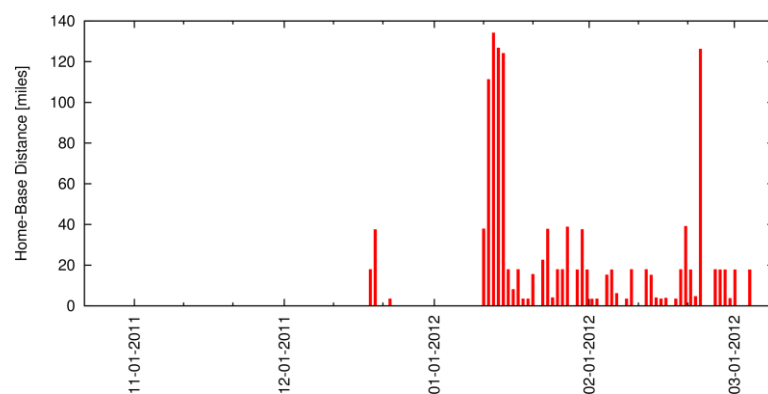


Figure 16 Maximum distance travelled from home-base for control subject Dafodil (Age 42)

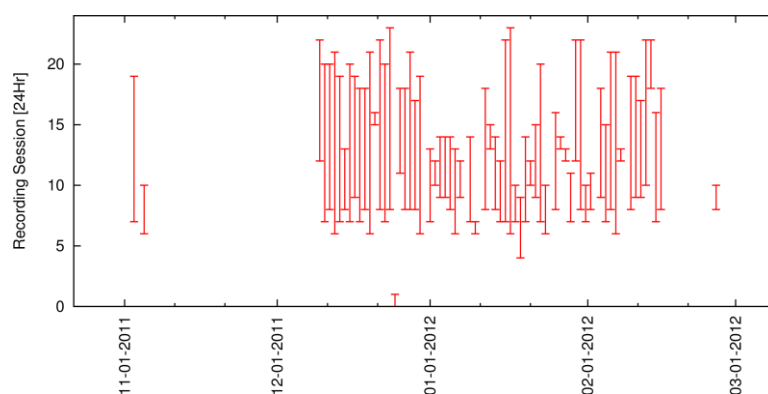


Figure 17 Start and End times for each recording session for diagnosed PD subject Daisy (Age 54; UPDRS 7)

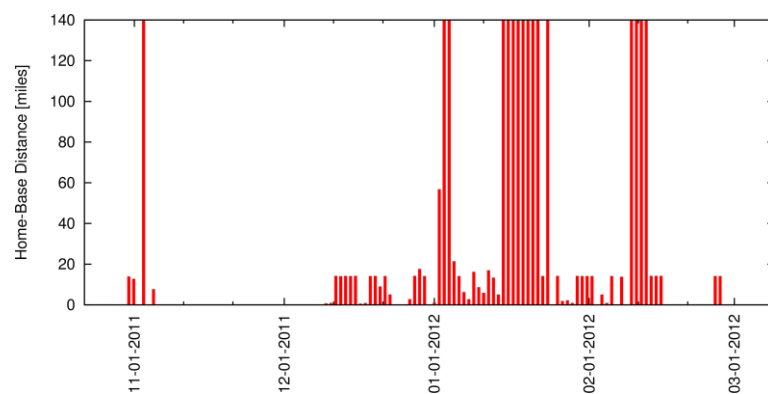


Figure 18 Maximum distance travelled from home-base for PD subject Daisy (Age 54; UPDRS 7)

Area (m²):

Desired Point Radius (m)

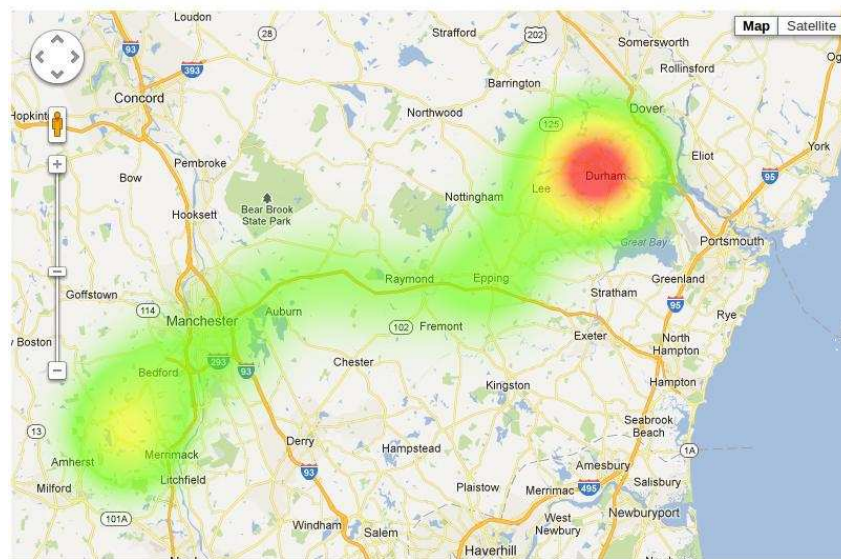


Figure 19 Heat-map of frequently visited locations for a control subject on a particular day with 10,000m resolution.

Area (m²):

Desired Point Radius (m)

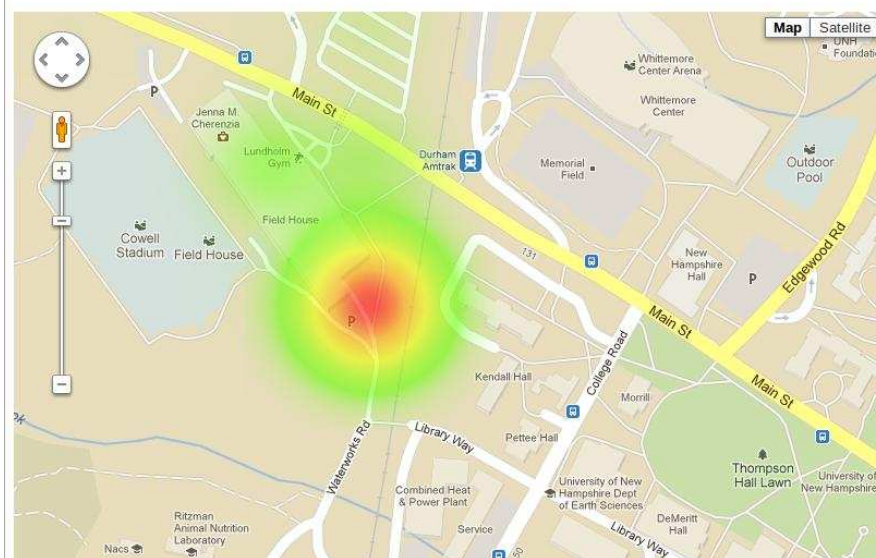


Figure 20 Heat-map of frequently visited locations for a control subject on a particular day with 100m resolution.

Area (m²):

Desired Point Radius (m)



Figure 21 Heat-map of frequently visited locations for a control subject on a particular day with 100 metre resolution. A enclosed polygon can be used to estimate the area of the subject occupied.

4 Conclusion

The MJFF's use of Kaggle to "crowd source" novel approaches to the analysis of the data set is significant because it brings the capabilities of smart-phones to the attention of a variety of stakeholders. In doing so the competition raises the level of stakeholder expectations; having seen what is possible from this competition, patients and funding bodies will demand more from future research studies and commercial applications. The industry sector and ultimately the patients suffering PD and other diseases will benefit from the improvements driven by the competition.

While the competition is likely to be of net positive benefit, a number of limitations are apparent in the study design:

- Small sample size
- Unusable data
- Privacy

The above issues will limit the scope of entries that MJFF can expect from competitors. For example, the quality of the accelerometer data mean entries that seek to correlate change in gait with PD progression may find their models of little value.

The AEHRC/APCN team brings an academically focussed, multi-disciplinary perspective to the competition. Our mapping of objective data acquired from sensors to clinical concepts such as life-space and activities of daily living will form the basis of a competitive entry. The team will gain valuable experience from the submission of an entry which will be useful in future collaborative projects.

5 Appendix A

Code Name	Classification	Gender	Current Age	Age when Diagnosed with Parkinson's Disease	UDPRS (pre)	UDPRS (post)	List any other chronic illnesses you are currently being treated for, or medications you are on
Apple	HC	Male	77	N/A	N/A	N/A	Irregular heartbeat medications..... low dose aspirin Coumadin 5 mg a day lisinopril 10 mg
Cherry	PD	Female	55	51	7	7	Had heart surgery 1 year ago to repair a mitral valve. medications are: Metaprolo 25mg Diovan 80mg baby aspirin 81mg Azilect 1mg Premarin 1.25mg Zocor 5mg Lovaza 1gram multi vitamin and vit c
Crocus	PD	Male	46	41	8	8	No other illnesses. Never been hospitalized. Never had surgery. Medications are for PD only - Azilect 1mg in morning; sinemet (sp) 1/2 pill 3 x daily; RequipXL 12mg one pill after dinner.
Dafodil	HC	Male	42	N/A	N/A	N/A	None supplied.
Daisy and Daisey	PD	Male	54	52	8	8	In addition to PD medications, I take Lipitor for cholesterol
Flox	PD	Male	57	47	23	23	Medications: Sinemet & Aricept Condition: Polycystic Kidney Disease Smoker Participating in a cognitive study with Dr. Press @ Beth Israel

Iris	PD	Male	65	45	16	17	Hydrocortisone Acetate Selegiline CCI Carbidopa-Levodopa-Entacapone Crbidopa-Levodopa Memantine Ropinirole Donepezil Simvastatin Trazodone Terazosin Omeprazole Vitamin D
Lilly and Lily	HC	Female	53	N/A	N/A	N/A	Graves Disease (Been treated, take medication now)
maple	PD	Male	55	46	26	23	sinemet 50/200 q3-4 hours;entacapone 200 mg with the sinemet. effexor xr 150 mg q am ativan 1mg 3/day prn tylenol #3 2 3/day prn Aspirin 325 1 qd artane 1-2 mg prn up to 6 mg/d amantadine 100mg qam rx for 1 q am and 1 at 12 pm[sleep late usually just 1 dose 12 pm seroquel 300 mg at 9 pm ativan 1 mg at 9 pm simvastatin 80 mg at 9 pm co-q10 b-vits homocysteine reduction n-acetyl cysteine otc papaya tabs-constipation
Orange	HC	Male	57	N/A	N/A	N/A	None supplied.
Orchid	PD	Male	69	65	15	10	High Blood Pressure (3 Medications) Sinemet, and an extended release form of Sinemet ComTam (Levedopa related) Prostate Cancer (In the watchful wait phase) Skin Cancer (Been treated a few times for it) Glockoma
Peony	PD	Male	80	67	13	14	Neuropathy
Rose	HC	Male	55	N/A	N/A	N/A	None supplied.

sunflower	HC	Male	67	N/A	N/A	N/A	diabetes high blood pressure
Sweetpea	HC	Female	77	N/A	N/A	N/A	hypo-thyroid medication .075 mg Synthroid once daily
violet	PD	Female	55	43 first symptoms, 53 first meds	5	5	No other illnesses 1 Azilect a day 1MG 2 Amantadine a day 1 in morning 1 at 1pm 100 MG an occ Diazepam 2.5 mg at night to sleep 2-4 200mg Advil a day

6 Glossary

Unified Parkinson's Disease Rating Scale [UPDRS] A rating scale used to follow the longitudinal course of Parkinson's disease. The UPDRS is the most commonly used scale in the clinical study of Parkinson's Disease

Hertz [Hz] A SI unit of frequency defined as the number of cycles per second produced by periodic phenomena.

Global Position Service [GPS] A space-based satellite navigation system that provides location and time information in all weather conditions, anywhere on or near the Earth where there is an unobstructed line of sight to four or more GPS satellites.

Power Spectrum Density [PSD] A function that shows the strength of the variations (energy) as a function of frequency.

Mel-frequency Cepstral [MFC] A representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

7 References

Boissy P, Briere S, Hamel M, Jog M, Speechly M, Karelis A, Frank J, Vincent C, Edwards R. Duval C and the EMAP group (2011) Wireless inertial measurement unit with GPS (WIMU-GPS) – Wearable monitoring platform for ecological assessment of lifespace and mobility in aging and disease. 33rd Annual IEEE International Conference of EMBS, Boston Massachusetts, USA, August 30 – September 3, 2011.

CONTACT US

t 1300 363 400
+61 3 9545 2176
e enquiries@csiro.au
w www.csiro.au

YOUR CSIRO

Australia is founding its future on science and innovation. Its national science agency, CSIRO, is a powerhouse of ideas, technologies and skills for building prosperity, growth, health and sustainability. It serves governments, industries, business and communities across the nation.

FOR FURTHER INFORMATION

ICT CENTRE – AUSTRALIAN E-HEALTH RESEARCH CENTRE

David Ireland
t +61 7 3253 3600
e david.ireland@csiro.au
w aehrc.com

ICT CENTRE – AUSTRALIAN E-HEALTH RESEARCH CENTRE

Simon McBride
t +61 7 3253 3631
e simon.mcbride@csiro.au
w aehrc.com