Predicting Parkinson's disease diagnostic with digital handwriting tests

Nina Caparros

2019-12-07

Contents

1 Introduction **3**
- I Nota bene . . . 3
- II Parkinson's disease . . . 3
- III Project overview . . . 4
  - A Material . . . 4
  - B Tests . . . 4
  - C Archimedean spiral . . . 5
  - D Goal . . . 6
- IV Dataset overview . . . 7

2 Method and analysis **8**
- I Initial Data . . . 8
- II Data cleaning . . . 9
  - A Format . . . 9
  - B Cleaning TimeStamp . . . 10
  - C Calibrating the samples . . . 11
- III Data analysis . . . 15
  - A Global analysis . . . 15
  - B Creating training and testing sets . . . 17
  - C Defining the Archimedean spiral . . . 17
  - D Distance (Static and Dynamic Spiral) analysis . . . 18
  - E Stability Test on Certain Point analysis . . . 19
- IV Issues . . . 20
  - A Tests inconsistency . . . 20
  - B Material inconsistency . . . 22
  - C Missing software and informations . . . 22
- V Prediction algorithms . . . 23
  - A Final prediction model . . . 23
  - B Chosing the prediction algorithms . . . 23
  - C Completing the training and testing sets . . . 23
  - D Parameters . . . 23
  - E Fitting models . . . 24

3 Results **25**

4 Conclusion **26**

5 Sources and references **27**

1 Introduction

This report presents the analysis and results of the "Choose Your Own Project" from the HarvardX's ninth course of the Data Science Professional Certificate Program available on edx.org. The chosen thematic was the prediction of the Parkinson's disease diagnosis depending on the results of three tests, measuring the motor performance, the tremor and the hand stability.

I Nota bene

The following section is a quick presentation of the Parkinson's disease but is not mandatory to understand this report.

When not relevant, the code used to create this report is run but not displayed. The complete source code can be found on GitHub (https://github.com/ncaparros/ParkinsonDisease_SpiralDrawing).

II Parkinson's disease

Parkinson's disease, sometimes abreviated as PD, is a long-term neurodegenerative disorder. Its cause is unknown, though it is believed to involve genetic (as relatives tend to contract the disease), and/or environmental factors (as pesticides).

The disease affects mostly the motor system, as tremor, akinesia (loss of the power of voluntary movement), shaking, rigidity, slowness of movement, difficulty with walking,... and as it worsen, it can cause depression, anxiety (more than a third of people with Parkinson's disease), emotional and sleep troubles, and in the advanced stages the disease can lead to dementia.

The motor symptoms of the Parkinson's disease (parkinsonian syndrome) are caused by the death of cells, more precisely dopaminergic neurons, in the *substantia nigra* (a region of the midbrain, see Figure 1, left). The *substantia nigra* is a structure divided into to parts : the *pars reticula* and the *pars compacta* (see Figure 1, left). It is the part of the brain that plays an important role in reward-seeking, learning and movement.

Dopamine is an organic chemical functioning as both an hormone and a neurotransmitter. Basically, neurotransmitters are chemical messengers which transmit signals by being released from one neuron to a receptor on the target cells. Neurotransmitters are critical to execute everyday functions as, in our case, movement (contact between a motor neuron and a muscle fiber).

The lack of dopamine (due to the death of those cells, and therefore induces a smaller substancia nigra than on a healthy subject, see Figure 1, right) provokes emotional troubles, and since the downsized *substancia nigra* is connected to the motor cortex (via the *pars reticula*, see Figure 1, left), it causes the parkisonian syndromes.

The Figure 1 (left) shows a lateral cross-section of a brain. The red arrows represent the dopamine's exchanges between the *pars reticula*.

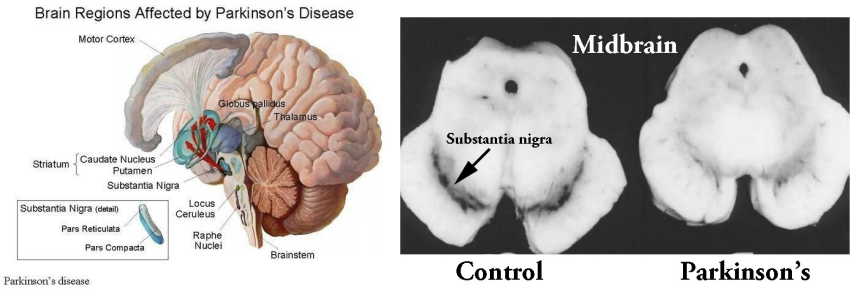

Figure 1: Lateral cross-section of the brain (left, source : http://www.neuroconvention.com/), *Substancia nigra* differences between a healthy brain and a Parkinson's brain (right, source : https://scienceofparkinsons.com/)

The Figure 1 (right) shows the lack of dopaminergic neurons in a brain of a person affected by Parkinson's disease compared to a healthy brain.

Parkinson's disease affected around 6.2 million people in 2015 and resulted in more than 117,000 deaths. This condition mostly occurs in people over the age of 60 (about one percent are affected). The average life expectancy following diagnosis is between 7 and 15 years.

III Project overview

In 2011, the Department of Neurology in Cerrahpasa Faculty of Medicine in Istanbul University (Turkey) provided a data set of test results from 62 patients with the Parkinson's disease and 15 from healthy people for a study (Muhammed Erdem Isenkul, Betul Erdogdu Sakar and Olcay Kursun) which purpose was to monitor Parkinson's disease with digitalized graphic tablets. The goal of this study was to provide easy access to Parkinson's disease progress monitoring to the elderly patients, or patients with an advanced stage of the disease, instead of the inconvenient and time-consuming process at the clinic. The tests aim to be non-invasive, would not require brain scans, would ease the work of the medical doctors, and would not require trained medical staff assigned to this task.

A Material

It was decided to perform three handwriting tests on a graphic tablet (Wacom Cintiq 12WX graphics, see Figure 2). The tablet would measure several parameters as : the coordinates (x-y-z) of the pen on the tablet, the pressure over the screen, the grip angle on the pen at regular time intervals. A software was developped in order to test the coordination of the patient.

B Tests

The three tests performed by the patients were :

Figure 2: Wacom Cintiq 12WX graphics, source : https://www.bhphotovideo.com/

- **Static Spiral Test (SST)**, a traditionnal test usually performed with paper and pencil. An Archimedean spiral (see Figure 3 and C *Archimedean spiral*) is printed on it, and the patient needs to retrace it. The more the patient suffer from an advanced stage of the Parkinson's disease, the more differences between the archimedean spiral and his drawing.
- **Dynamic Spiral Test (DST)**, a new test introduced in the study, where the archimedean spiral *blinks*. It is only seen at certain times. It becomes more difficult to follow the spiral.
- **Stability On Certain Point Test (SOCPT)**, a test where there is a red point in the middle of the tablet's screen, and the patients are asked to hold the pen on the point without touching the screen. This test determines the patient's hand stability and hand tremor level.

C Archimedean spiral

The Archimedean spiral (named after the Greek mathematician Archimedes), can be described by the polar coordinates equation :

$$r = a + b\theta$$

With r the radius, a real number turning the spiral, b real number controling the distance between successive turns, and θ the angle velocity. The spiral is the locus of points corresponding to the locations over time of a point moving away from a fixed point with constant speed along a line that rotates with constant angular velocity.

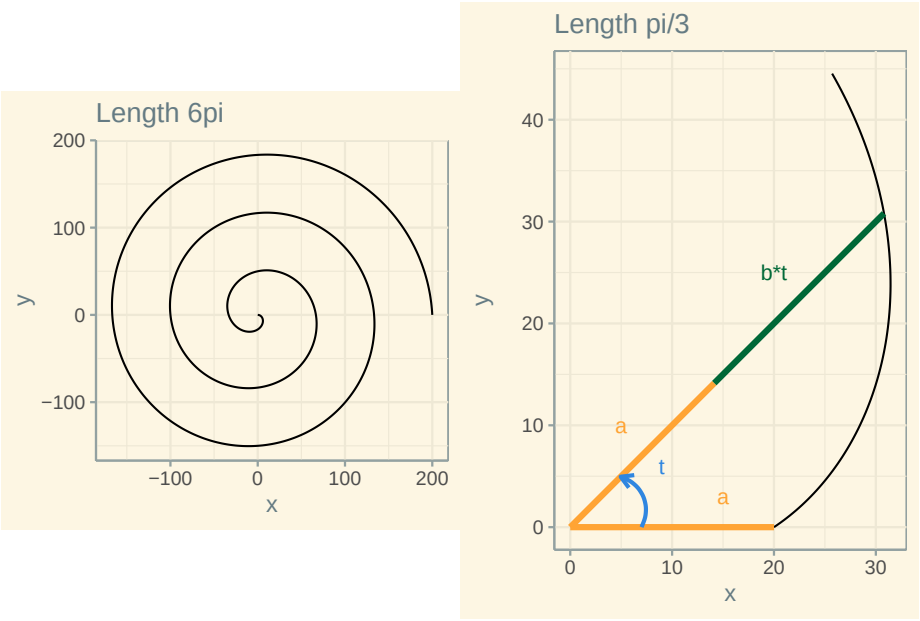

Figure 3: Archimedean spirals

It can be described in the cartesian coordinates by :

$$x = (a + bt) * cos(t)$$

$$y = (a + bt) * sin(t)$$

The Figure 3 (left), shows an Archimedean spiral similar to the one used in the SST and DST. Here, $a = 0$, $b = 10.61$, and the length is 6π (t ranges from 0 to 6π). This means that the spiral does three complete turns.

The Figure 3 (right), shows an Archimedean spiral with a length of only $\frac{\pi}{3}$, to show the different parameters. Here, $a = 20$ and $b = 30$, and we have the following equations :

$$\begin{cases} x_b(t) = bt * cos(t) \\ y_b(t) = bt * sin(t) \\ b(t) = \sqrt{x_b(t)^2 + y_b(t)^2} = \sqrt{(bt * cos(t))^2 + (bt * sin(t))^2} \\ = \sqrt{b^2t^2(cos^2(t) + sin^2(t))} = \sqrt{b^2t^2} = bt \\ b(t = 0) = b * t = 0 \end{cases}$$

D Goal

In this report I will build an algorithm able to predict if the patient has or has not Parkinson's disease based on this dataset. I do not have access to the software used in the study, so I will have to recreate and approximate the Archimedean

spiral. I do not have access to the *scores* of the patients, given by neurologists, representing the stage of Parkinson's disease. Therefore, my output will only be a boolean, has, or has not, with at best a percentage of probability, not a neurological scale. If efficient, these results would help determine if a person taking the test should consult a specialist for further examination. The original study based its conclusion of the acceleration (change of velocity) for the firsts two tests (static and dynamic spiral tests). I will try instead to use the areas and lengths of each drawing, as well as the difference of time needed to perform the two firsts tests.

IV Dataset overview

The dataset provided was an archive .zip containing three folders. One of them was composed by only .png images of the tests results, which were already saved in the text dataset. In the two remaining folders, there were datasets related to healthy (called controls) and people with Parkinson's disease (called PWP, People With Parkinson). Since the datasets in both folders were following the same pattern, they were merged in a single one.

Each text file of the dataset was the test results of a single patient. Each line of the file represented one measure, at a certain time, of the X-Y-Z coordinates of the digital pen, the pressure of the pen on the screen, the grip angle, the timestamp (at which the measure had been taken) and the test identifier (Static Spiral Test : 0, Dynamic Spiral Test : 1, Stability Test on Certain Point Test :2).

The data was presented as *X;Y;Z;Pressure;GripAngle;Timestamp;TestID*:

191;205;0;39;1350;17535179;0

191;205;0;54;1360;17535186;0

191;205;0;60;1350;17535193;0

191;205;0;61;1360;17535200;0

2 Method and analysis

In this section I describe the process, from analysing the original dataset, to building the prediction algorithms, through data cleaning and analysis.

I Initial Data

Once the data downloaded and the data frame built (see IV *Dataset overview*), the first step is to add a random identifier to each patient and to note if he has Parkinson's disease or not. The random identifier has been chosen because :

- Since the text files were from different folders, and as some files had the same numbers, no pattern could be used for identifiers.
- It allows to not get focused on the patient id.

V1	patientID	isPwp
200;204;0;73;910;1732647300;0	3984365604	FALSE
200;204;0;218;900;1732647307;0	3984365604	FALSE
200;204;0;253;900;1732647314;0	3984365604	FALSE
200;204;0;304;900;1732647321;0	3984365604	FALSE
200;204;0;351;900;1732647328;0	3984365604	FALSE
200;204;0;386;900;1732647335;0	3984365604	FALSE

Then, each value has to be extracted into a new column of the data frame.

X	Y	Z	Pressure	GripAngle	Timestamp	TestID	patientID	isPwp
200	204	0	73	910	1732647300	0	3984365604	FALSE
200	204	0	218	900	1732647307	0	3984365604	FALSE
200	204	0	253	900	1732647314	0	3984365604	FALSE
200	204	0	304	900	1732647321	0	3984365604	FALSE
200	204	0	351	900	1732647328	0	3984365604	FALSE
200	204	0	386	900	1732647335	0	3984365604	FALSE

X and Y represents the place of the pen on the tablet, we can assume horizontally and vertically, and Z is the height between the pen and the screen. A Z equal to 0 means the pen is on the screen.

A new data frame containing one line by patient is then created and filled as :

patientID	isPwp
3984365604	FALSE
3504064925	FALSE
2537088573	FALSE

patientID	isPwp
3057315719	FALSE
1117875664	FALSE
3882243293	FALSE

It will be used for summaries and additionnal informations on the patient or the test later.

II Data cleaning

In this subsection I explain the process of formating the values, cleaning the Timestamp column and calibrating the test samples.

A Format

Since the datas were extracted from a text file, they were all, but the two we added, of class character.

```
#Shows the class of each column in the df data frame
lapply(df, class)
```

```
## $X
## [1] "character"
##
## $Y
## [1] "character"
##
## $Z
## [1] "character"
##
## $Pressure
## [1] "character"
##
## $GripAngle
## [1] "character"
##
## $Timestamp
## [1] "character"
##
## $TestID
## [1] "character"
##
## $patientID
## [1] "numeric"
##
```

```
## $isPwp
## [1] "logical"
```

It was impossible then to perform any action on those values, so they are all to be converted as numeric with the `as.numeric` function.

B Cleaning TimeStamp

The timestamp columns seems pretty obscur, and trying to parse it into a readable date would produce either an impossible date (`make_date` or `make_datetime`) or NA values (`dym`, `mdy_hms`, `as.Date`,...).

```
#The date time functions would not give any satisfying result
make_date(as.character(df[1,]$Timestamp))
```

```
## [1] "-5877641-06-23"
```

```
make_datetime(df[1,]$Timestamp)
```

```
## [1] "1732647300-01-01 UTC"
```

```
dym(as.character(df[1,]$Timestamp))
```

```
## Warning: All formats failed to parse. No formats found.
```

```
## [1] NA
```

```
as.Date(as.character(df[1,]$Timestamp), "%Y%M%D")
```

```
## [1] NA
```

To be able to use the timestamp more easily, and mostly because we do not know its unit (probably milliseconds but we cannot know for sure), I substract the first timestamp of every couple test/patient to all the timestamp values, making the first value 0. To do this I have to create a new empty data frame, and two for loops : one for the tests (0 to 2) and one for the patients (0 to 77). Inside the loops, I would get the values of the current patient for the current test, arranged by ascending timestamp, and the first value would be the initial timestamp value. Then, every timestamp would be mutated as $timestamp_i = timestamp_i - timestamp_0$. The mutated data frame would then be merge (`rbind`) into the final data frame.

X	Y	Z	Pressure	GripAngle	Timestamp	TestID	patientID	isPwp
200	204	0	73	910	1732647300	0	3984365604	FALSE
200	204	0	218	900	1732647307	0	3984365604	FALSE
200	204	0	253	900	1732647314	0	3984365604	FALSE
200	204	0	304	900	1732647321	0	3984365604	FALSE
200	204	0	351	900	1732647328	0	3984365604	FALSE
200	204	0	386	900	1732647335	0	3984365604	FALSE

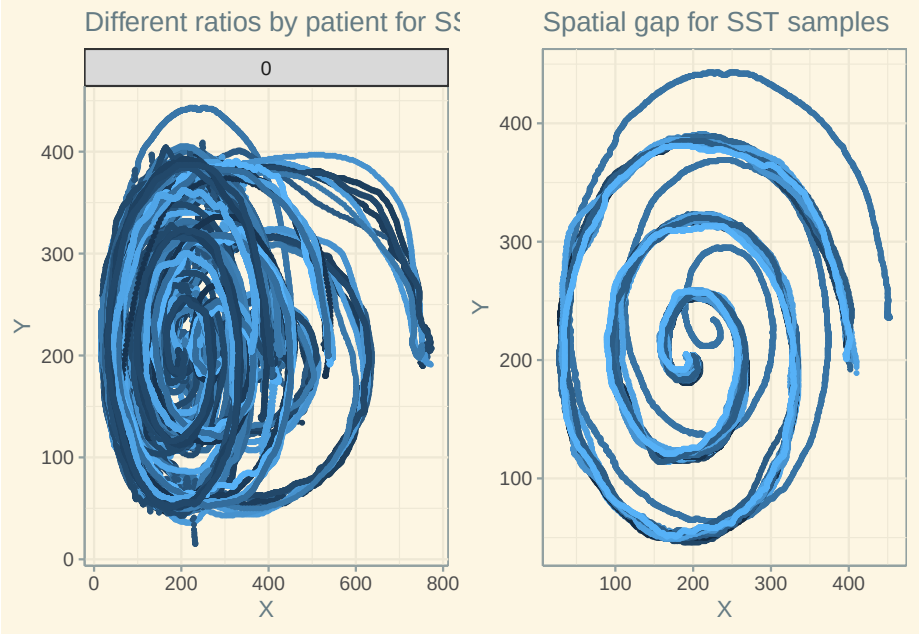

Figure 4: SST samples

X	Y	Z	Pressure	GripAngle	Timestamp	TestID	patientID	isPwp
200	204	0	73	910	0	0	3984365604	FALSE
200	204	0	218	900	7	0	3984365604	FALSE
200	204	0	253	900	14	0	3984365604	FALSE
200	204	0	304	900	21	0	3984365604	FALSE
200	204	0	351	900	28	0	3984365604	FALSE
200	204	0	386	900	35	0	3984365604	FALSE

With this modification, it is easier to compare the timestamps between patients and tests, since now the first value of every test by a patient is 0.

C Calibrating the samples

A quick look at the test results (the drawings) shows that the samples are not all consistent in term of X-Y ratio (see Plot 4, left). Seven samples can be seen with a X-Y ratio of almost 2 instead of 1. With the Archimean spiral formula in mind, it seems odd that a small part of the patients would have been given a different test with a different drawing to follow.

Another odd detail is a spacial gap between patterns with the same ratio (see Plot 4, right). One of the control sample seems to have been shifted.

This disparity between the samples prevented me from trying to analyse the spacial differences between the generated Archimedean spiral and the patient's

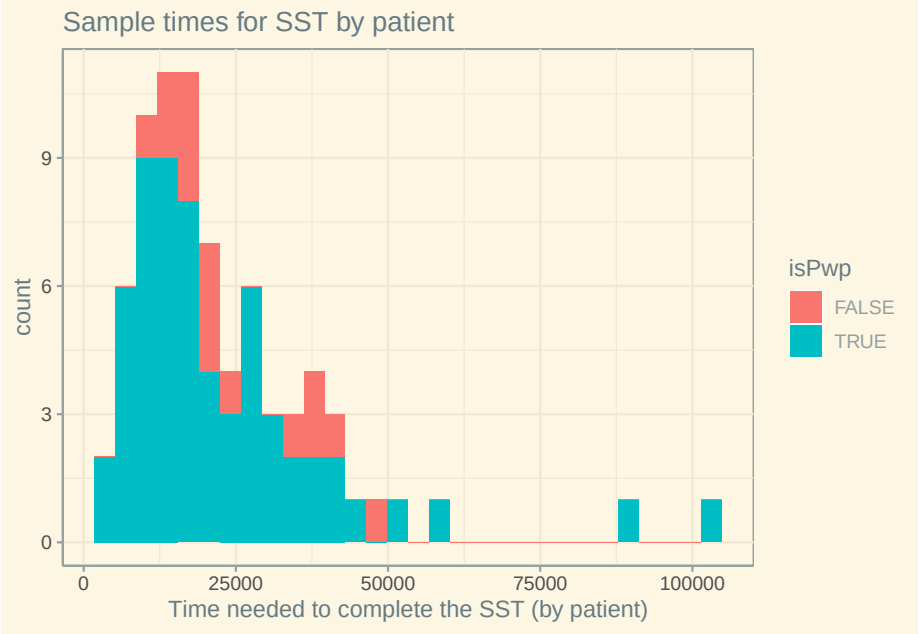

Figure 5: SST sample times

drawing. I checked if the timestamp data seemed affected or not (see Plot 5). It seemed as only the X and Y values were sometimes off.

To calibrate the X and Y values, the ratios $\frac{X}{Y}$ are calculated and stored in a new data frame *ratiosDf*, by patient and test. This data frame contains the ratio between X and Y as

$$ratio = \frac{X_{max} - X_{min}}{Y_{max} - Y_{min}}$$

We can indeed see that even if most ratios are around 1.3, some outliers are around 2 (see Figure 6).

Even the initial point of the drawing is different (see Figure 7).

To standardize all the spirals, several parameters are taken into account :

- The following ratio should be of one :

$$\frac{X_{max} - X_{min}}{Y_{max} - Y_{min}}$$

- The height and length of the spiral should be the same (approximately)
- The X and Y values would have to be slid

Standardizing the ratio would simply be

$$X_{standardized_i} = \frac{X_{old_i}}{ratio}$$

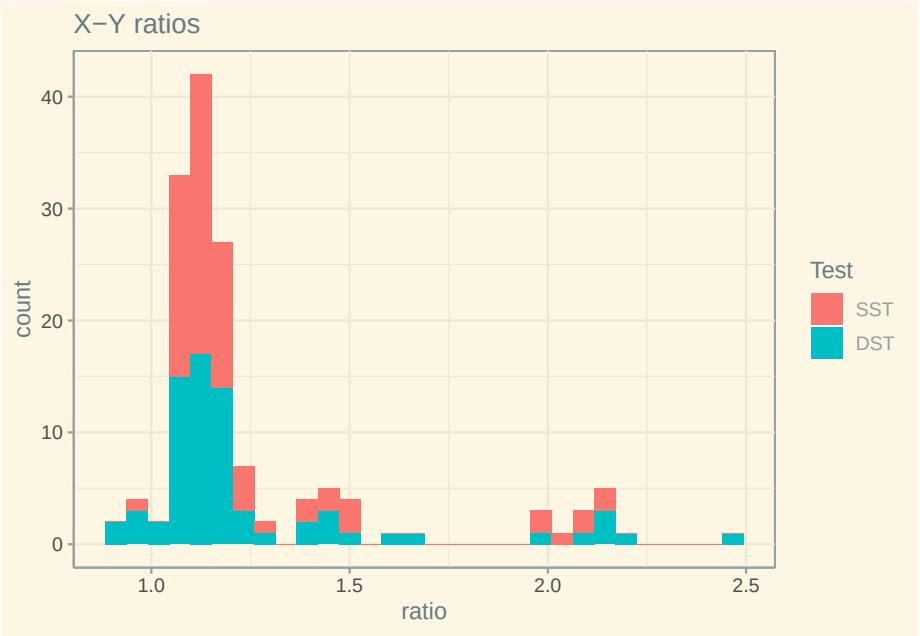

Figure 6: X-Y ratios before cleaning

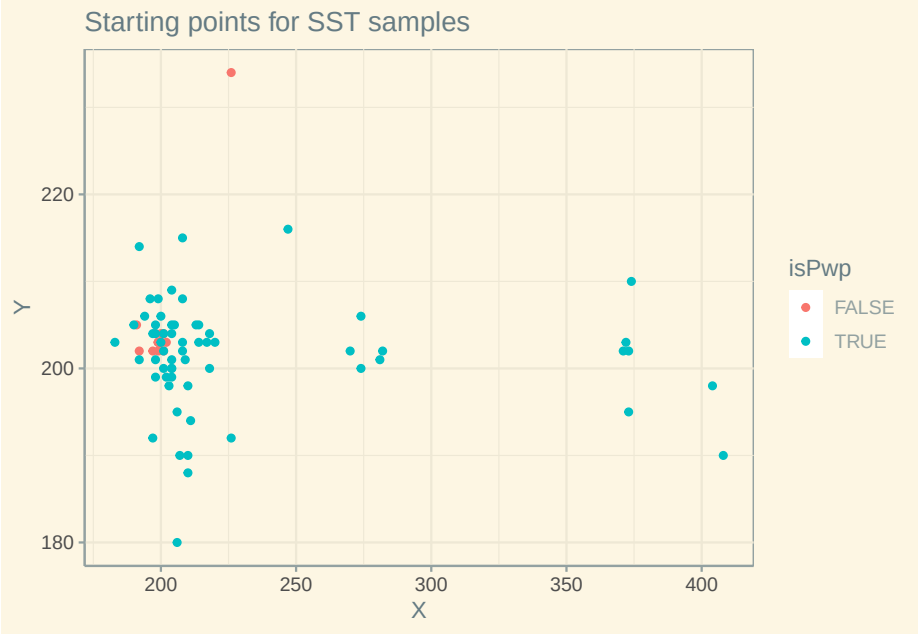

Figure 7: SST starting point

and this would imply that

$$\frac{X_{maxStandardized} - X_{minStandardized}}{Y_{max} - Y_{min}} = 1$$

.

Resizing the height and length of the spiral would be done with :

$$X_{resized_i} = \frac{X_{old_i}}{Spiral_{length}} * Size$$

and

$$Y_{resized_i} = \frac{Y_{old_i}}{Spiral_{length}} * Size$$

when $Spiral_{length} = Spiral_{height}$ since the ratio

$$\frac{X_{maxStandardized} - X_{minStandardized}}{Y_{max} - Y_{min}} = 1$$

as seen previously, and with $Size$ is the desired size of the spiral. I chose an arbitrary $Size = 400$ and since the unit is not given I assume it is $400px$. The unit would not matter in the study.

In order to replace the spirals near the origin of the graph ($X = 0$ and $Y = 0$), two adjustment parameters need to be determined. The formula would simply be $X_{slid_i} = X_{old_i} - X_{param}$ and $Y_{slid_i} = Y_{old_i} - Y_{param}$.

Several methods were tried for determining the best X_{param} and Y_{param} such as:

- $X_{param} = \overline{X}$ and $Y_{param} = \overline{Y}$
- $X_{param} = X_{median}$ and $Y_{param} = Y_{median}$
- $X_{param} = \overline{X_{Y=Y_{median}}}$ and $Y_{param} = \overline{Y_{X=X_{median}}}$
- $X_{param} = \overline{X_{Y=Y_{mean}}}$ and $Y_{param} = \overline{Y_{X=X_{mean}}}$
- $X_{param} = \overline{X_{Y=Y_{Timestamp=0}}}$ and $Y_{param} = \overline{Y_{X=X_{Timestamp=0}}}$
- $X_{param} = X_{initial}$ and $Y_{param} = Y_{initial}$

And eventually the most accurate models are :

$$X_{param} = \frac{(X_{initial} + X_{min} + X_{max})}{3}$$

$$Y_{param} = \frac{(Y_{initial} + Y_{min} + Y_{max})}{3}$$

This model takes into account the assumption of which the very first values X,Y are, in general, the closest to the Archimedean Spiral when $Timestamp = 0$. X_{min} and X_{max} (and respectively Y_{min} and Y_{max}) are approximations of the very tip of the drawings (see Figure 8).

Combining those three models gives for patient p, test t and spiral length L equal spiral height (see Figure 9):

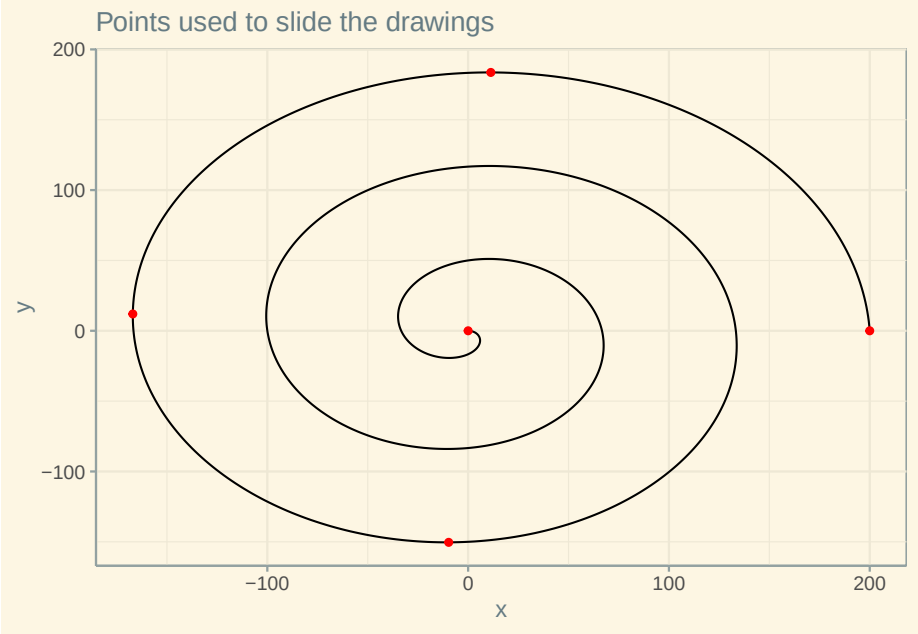

Figure 8: Archimedean spiral

$$X_{i,p,t} = \frac{\frac{X_{old_i} - X_{param}}{ratio_{p,t}}}{L_{p,t}} * 400 = \frac{(X_{old_i} - X_{param}) * 400}{ratio_{p,t} * L_{p,t}} = 400 \frac{(X_{old_i} - \frac{X_{initial} + X_{min} + X_{max}}{3})}{ratio_{p,t} * L_{p,t}}$$

$$Y_{i,p,t} = \frac{Y_{old_i} - Y_{param}}{L_{p,t}} * 400 = 400 \frac{(Y_{old_i} - \frac{Y_{initial} + Y_{min} + Y_{max}}{3})}{L_{p,t}}$$

III Data analysis

This section describes the analysis of the different tests in the dataset. Since there is no prediction here, and some additionnal information (calculated and determined values) are to be noted in the `patients` data frame, the analysis is performed on the whole datas.

A Global analysis

The very first point to stress out is that every patient did not perform the three tests (see Figure 10).

This would probably alter the accuracy of the future predictions. The original study based its analysis on the combination of SST and DST, and even though every patient took the SST (Static Spiral Test), some DST (Dynamic Spiral Test)

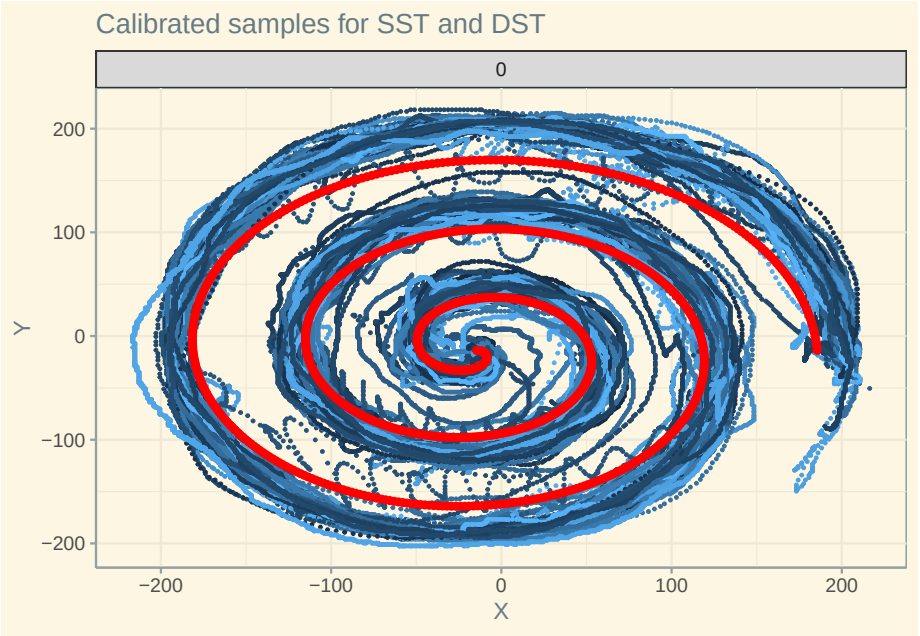

Figure 9: Calibrated SST and DST samples

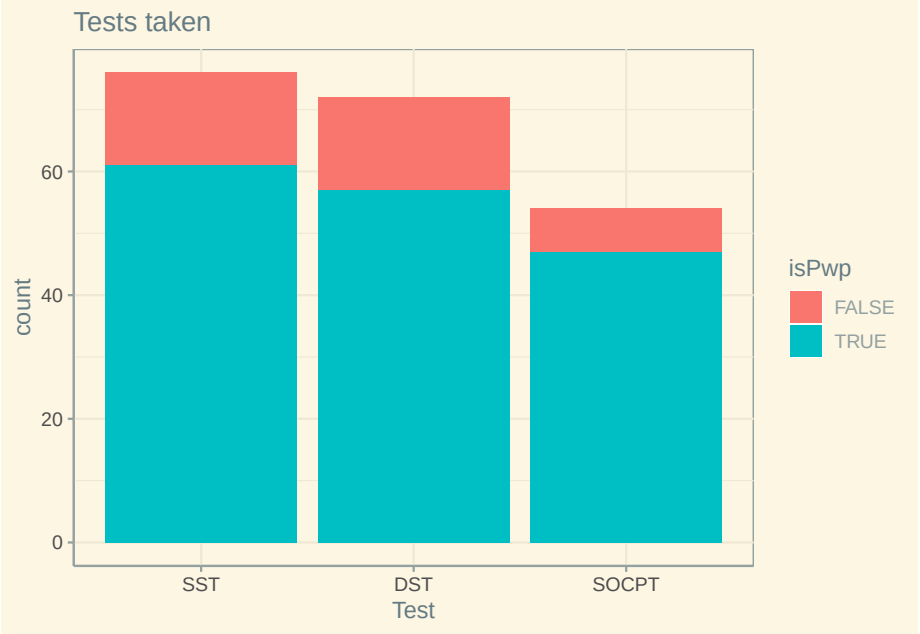

Figure 10: Tests taken

were missing from the dataset. This issue is detailed and handled on section *Test inconsistency* (see A).

B Creating training and testing sets

In order to calibrate the Archimedean spiral, I need to create training and testing sets. Since we have very little values, I need to make sure I have control patients in both sets, as well as patients (control and not) who had taken the Stability on Certain Point Test. I create several sub-datasets as :

- `subControls` : a dataset with only control patients, who had not taken the Stability on Certain Point Test
- `subParkinsons`: a dataset with only patients with the Parkinson's disease, who had not taken the Stability on Certain Point Test
- `subTest2Ctrl` : a dataset with only control patients who had taken the Stability on Certain Point Test
- `subTest2Parkinsons` : a dataset with only patients with the Parkinson's disease, who had taken the Stability on Certain Point Test

This method is way less simple than the `createDataPartition` function, but allows me to get a sample of each case in both training and testing sets.

C Defining the Archimedean spiral

The equation of the Archimedean spiral can be described with :

$$\begin{cases} x(t) = a + bt * cos(t) \\ y(t) = a + bt * sin(t) \end{cases}$$

I define $a = 0$ and determine that the spiral ranged from 0 (the origin of the spiral) to 6π (the last point of the spiral). This means that the spiral did three complete turns. I had to make a slight modification so the Archimedean spiral would match the drawings, as :

$$\begin{cases} x(t) = a + bt * cos(-t) = a + bt * cos(t) \\ y(t) = a + bt * sin(-t) = a - bt * sin(t) \end{cases}$$

I fix the maximum size of the spiral to 200 (see C *Calibrating the samples*), so I have the equation :

$$\begin{cases} x(t_{max} = 6\pi) = bt * cos(6\pi) = bt = b * 6\pi = 200 \\ y(t_{max}) = -bt * sin(6\pi) = 0 \end{cases}$$

Which means that $b = \frac{200}{6\pi} \approx 10.61$. I also need to slide the spiral, with $x_{adjust} = \overline{x(t = 0, isPwp = FALSE)}$ and $y_{adjust} = \overline{x(t = 0, isPwp = FALSE)}$ so that it is overlapping the drawings. The equation is then :

$$\begin{cases} x(t) = 10.61 * t * cos(t) + x_{adjust} \\ y(t) = -10.61 * t * sin(t) + y_{adjust} \end{cases}$$

D Distance (Static and Dynamic Spiral) analysis

I build a method which would estimate the difference between the Archimedean Spiral and the patient's drawing to be used for the analysis of the first two tests, using the `rgeos` package and the notion of `Spatial points` to work with spatial coordinates.

The logic behind the algorithm is quite simple. I define four variables in the `patients` dataframe :

- `areaT0` (the area between the drawing of the SST and the Archimedean spiral)
- `areaT1` (the area between the drawing of the DST and the Archimedean spiral)
- `sdT0` (the standard deviation between the SST and the Archimedean spiral)
- `sdT1` (the standard deviation between the DST and the Archimedean spiral)

For each unique point on the drawing (SST and DST), I determine the closest point on the Archimedean spiral using `gDistance` of the `rgeos` package and `which.min`. I then calculate the distance using :

$$d_i = \sqrt{(x_{test} - x_{spiral})^2 + (y_{test} - y_{spiral})^2}$$

With d_i the distance of the point $i = (x_{test}, y_{test})$ of the drawing and it's closest point on the Archimedean spiral (x_{spiral}, y_{spiral}). The area between the drawing and the spiral is approximated by :

$$A_{T_i} = \sum_i d_i$$

This approach doed not seem very conclusive (see Figure 11). Even if the results of the healthy patient are close, many patients with Parkinson's disease have similar results. It seems more appropriate to compare instead the drawings of the first two tests and to compute their area difference (`areaT0_T1`), and standard deviation (`sdT0_T1`). I also compute the time difference (`difft1_t0`) between those two tests. The algorithm is similar to the one used for the comparison of the drawings and the Archimedean spiral.

For each unique point on the SST drawing, I determine the closest point on the DST drawing using `gDistance` of the `rgeos` package and `which.min`. I then calculate the distance using :

$$d_i = \sqrt{(x_{SST} - x_{DST})^2 + (y_{SST} - y_{DST})^2}$$

With d_i the distance of the point $i = (x_{SST}, y_{SST})$ of the SST drawing and it's closest point on the DST drawing (x_{DST}, y_{DST}). The area between the drawing and the spiral is approximated by :

$$A_{T_i} = \sum_i d_i$$

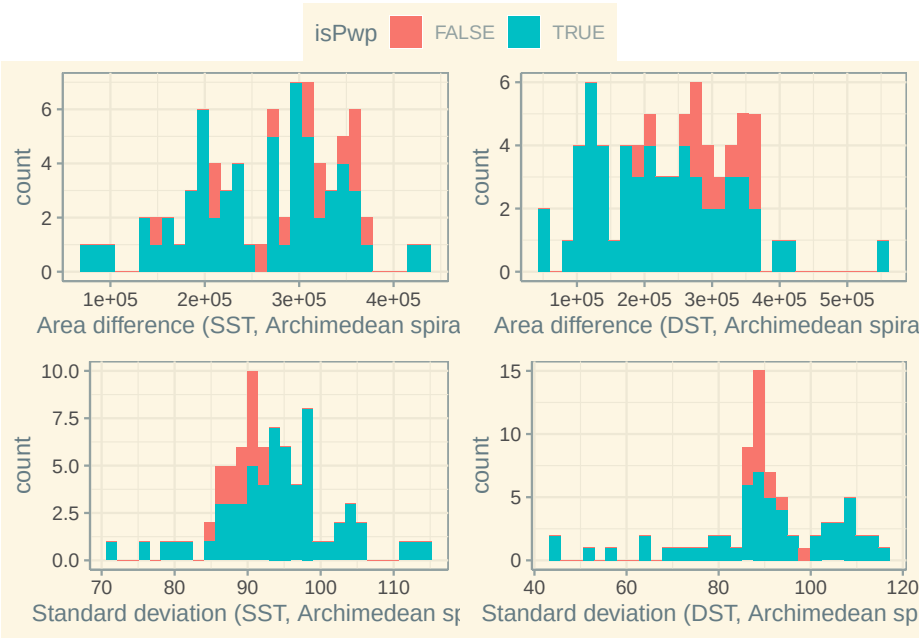

Figure 11: SST - Archimedean spiral comparison

The time difference is :

$$\Delta = t_{DSTmax} - t_{SSTmax}$$

With this method, comparing the Static Spiral Test and the Dynamic Spiral Test, it is easier to categorize the patients (see Figure 12). The results are more distinct.

E Stability Test on Certain Point analysis

This test would give important information about the hand tremor of the patient. If the pen is to touch the screen, or move a lot over the tablet, the patient would probably have the Parkinson's disease. To quantify this, I compute the number of points on the screen (basically every point where Z equal to 0) and the total distance on the X and Y axis of the pen. The distance between two points is :

$$d_i = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}$$

The total distance of the pen over the tablet is :

$$D = \sum_i d_i$$

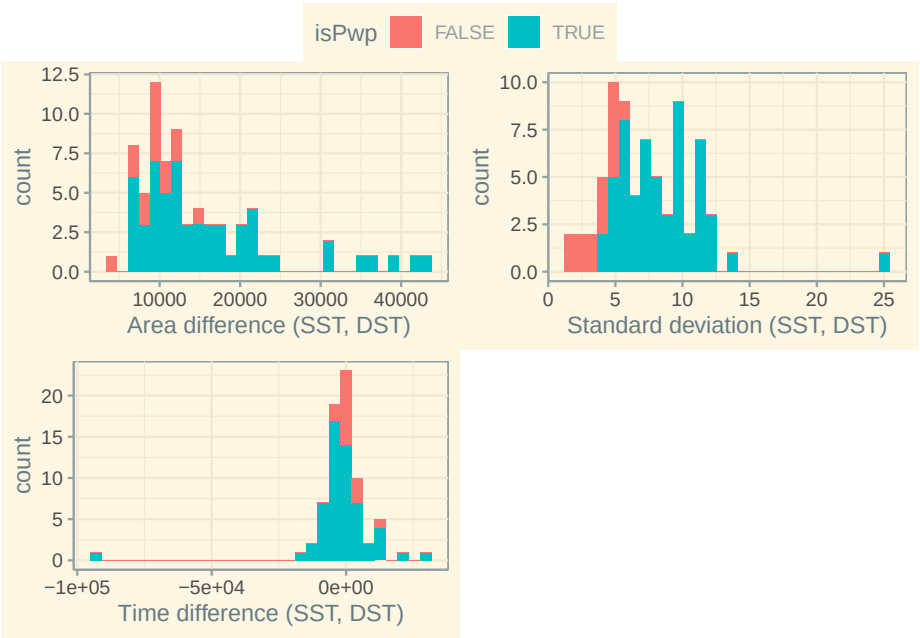

Figure 12: SST-DST comparison

IV Issues

A Tests inconsistency

The very first issue was that every patient did not take the three tests. I needed then to come up with algorithms managing the different cases.

```
#build the data frame with patients and tests
couplesTestsPatients <- df %>%
  select(patientID, TestID) %>%
  group_by(patientID, TestID) %>%
  unique()

#get the patients who had taken the SST
PatientsT0 <- couplesTestsPatients %>%
  filter(TestID == 0)

#get the patients who had taken the DST
PatientsT1 <- couplesTestsPatients %>%
  filter(TestID == 1)

#get the patients who had taken the SOCPT
PatientsT2 <- couplesTestsPatients %>%
```

```
  filter(TestID == 2)

#Are there any patients who took the SST
#but neither the DST nor the SOCP?
PatientsT0 %>%
  filter(!patientID %in% PatientsT1$patientID &
           !patientID %in% PatientsT2$patientID) %>%
  nrow() > 0
```

```
## [1] FALSE
```

```
#Are there any patients who took the SST
#and the DST but not the SOCP?
PatientsT0 %>%
  filter(patientID %in% PatientsT1$patientID &
           !patientID %in% PatientsT2$patientID) %>%
  nrow() > 0
```

```
## [1] TRUE
```

```
#Are there any patients who took the SST
#and the SOCP but not the DST?
PatientsT0 %>%
  filter(!patientID %in% PatientsT1$patientID &
           patientID %in% PatientsT2$patientID) %>%
  nrow() > 0
```

```
## [1] TRUE
```

```
#Are there any patients who took the SST, DST and SOCP ?
PatientsT0 %>%
  filter(patientID %in% PatientsT1$patientID &
           patientID %in% PatientsT2$patientID) %>%
  nrow() > 0
```

```
## [1] TRUE
```

```
#Are there any patients who took the DST
#but neither the SST nor the SOCPI?
PatientsT1 %>%
  filter(!patientID %in% PatientsT0$patientID &
           !patientID %in% PatientsT2$patientID) %>%
  nrow() > 0
```

```
## [1] FALSE
```

```r
#Are there any patients who took the DST
#and the SOCP but not the SST?
PatientsT1 %>%
  filter(patientID %in% PatientsT2$patientID &
           !patientID %in% PatientsT0$patientID) %>%
  nrow() > 0
```

```
## [1] FALSE
```

```r
PatientsT2 %>%
  filter(patientID %in% PatientsT1$patientID &
           !patientID %in% PatientsT0$patientID) %>%
  nrow() > 0
```

```
## [1] FALSE
```

```r
#Are there any patients who took the SOCP
#but neither the SST nor the DST?
PatientsT2 %>%
  filter(!patientID %in% PatientsT0$patientID &
           !patientID %in% PatientsT1$patientID) %>%
  nrow() > 0
```

```
## [1] TRUE
```

Thoses four different cases are patients, who :

- Took the Static Spiral Test and only one of the other test (Dynamic Spiral Test or Stability On Certain Point Test)
- Took all three tests
- Took only the Stability On Certain Point Test

B Material inconsistency

As seen in the *Data Cleaning - Calibrating the samples* (see II) section, the records were not on the same scale and sometimes not on the same starting point.

C Missing software and informations

Another issue was that I had no access to the software used to record the tests, nor to the configuration of the Archimedean spiral used for the tests. I did not have the frequency at which the DST blinked. My study and models were therefore limited.

V Prediction algorithms

In this section I define the predition models and different algorithms.

A Final prediction model

Based on the previous observations I define independent predictions as :

- Prediction on the Static Spiral Test compared to the Dynamic Spiral Test ($\hat{Y}_{SST\sim DST}$)
- Prediction on the Stability On Certain Point Test ($\hat{Y}_{SOCPT}$)

When it is not possible to do the prediction due to the tests recorded, the value would be `NA`. I do a weighted average, figuring that the prediction on the Stability On Certain Point Test is more precise than the prediction on the Static Spiral Test compared to the Dynamic Spiral Test as :

$$\hat{Y} = \overline{\hat{Y}_{SST\sim DST} + 4 * \hat{Y}_{SOCPT}}$$

When there is a missing prediction, it is removed from the equation.

B Chosing the prediction algorithms

I picked the random forest model (`randomForest`). The k-nearest neighboors and the linear regression were not approriate for the data gathered and computed. I would do the weighted mean between the two predictions.

$$\hat{Y} = \overline{\hat{Y}_{SST\sim DST,rf} + 4 * \hat{Y}_{SOCPT,rf}}$$

C Completing the training and testing sets

In order to use all the parameters calculated and computed, I need to merge the training and testing set with the complete `patients` data frame. This allow me to not create testing and training sets again with the complete datas.

```
trainingSet <- left_join(trainingSet, patients %>%
                           select(-isPwp), by = "patientID")

testingSet <- left_join(testingSet, patients %>%
                          select(-isPwp), by = "patientID")
```

D Parameters

The variables used for each prediction model are :

- Static Spiral Test compared to the Dynamic Spiral Test ($\hat{Y}_{SST\sim DST}$) : area difference between the SST and the DST (`sdT0_T1`), its standard deviation (`sdT0_T1`) and the difference of execution time between the two tests (`difft1_t0`)

- Stability On Certain Point Test ($\hat{Y}_{SOCPT}$) : the number of points where the pen touched the screen (`Z0_T2`) and the total distance on the plan (`X`,`Y`) of the pen (`DistT2`)

It is required to mutate `isPwp` to be a numeric instead of a boolean.

E Fitting models

It is time to define the six fitting models and compute the different predictions:

```
#Static Spiral Test compared to the Dynamic Spiral Test
#random forest
fit_SST_DST <- randomForest(isPwp ~ areaT0_T1 +
                              sdT0_T1 +
                              difft1_t0,
                               data = trainingSet,
                            na.action = na.omit)

yhat_SST_DST <- predict(fit_SST_DST,
                        testingSet)


#Stability On Certain Point Test
#random forest
fit_SOCPT <- randomForest(isPwp ~ Z0_T2 +
                            DistT2,
                             data = trainingSet,
                          na.action=na.omit)

yhat_SOCPT <- predict(fit_SOCPT,
                      testingSet)


#Computing the predictions
yhatDf <- data.frame(yhat_SST_DST,
                     yhat_SOCPT)

p_hat <- yhatDf %>%
  rowwise() %>%
  mutate(phat = weighted.mean(c(yhat_SST_DST,
                                yhat_SOCPT),
                              c(1, 4),
                              na.rm = TRUE)) %>%
  select(phat)
```

3 Results

```
#if the prediction is above 0.5,
#the patient is diagnosed as with Parkinson's
Yhat <- ifelse(p_hat > 0.5, 1, 0) %>%
  factor()

#binding the prediction,
#the diagnosis and the medical diagnosis
results <- cbind(p_hat, Yhat, testingSet$isPwp)
results
```

phat	Yhat	testingSet$isPwp
0.2662333	0	0
0.1491585	0	0
0.9269667	1	1
1.0000000	1	1
1.0000000	1	1
0.9717333	1	1
1.0000000	1	1
0.9649000	1	1
0.8294985	1	1

```
#get the accuracy
acc <- confusionMatrix(Yhat,
                testingSet$isPwp %>%
                  factor())$overall["Accuracy"]

acc
```

```
## Accuracy
##        1
```

The accuracy we got with this model is of 1, but, it should be treated cautiously. Since we do not have much data in the dataset, we have no certainty about the accuracy for the people with the disease at it's very first stage. It is also complicated to apprehend since we do not have access to the medical "score" for each patient, instead we have a binary result, healthy, or with Parkinson's. In the ideal scenario, we would have had access to a scale, with the "progress" of the disease (which stage at least).

4 Conclusion

The original study shows that it is possible to acquire handwriting tests and monitor Parkinson's disease with a digital tablet, using mostly the difference of acceleration for the Static Spiral drawing and the Dynamic Spiral drawing.

I showed here that it was also possible to compute the area difference between those two tests, as well as the distance traveled by the pen on the Stability On Certain Point test, and that feeding these results in a random forest algorithm could predict accurately if the patient had or not Parkinson's disease.

These tests could be used to determine if a patient should proceed with the medical tests and consult a specialist, without having to go to an hospital and get the invasive neurological tests (if not needed). When using `phat` (the prediction of the random forest algorithms), it could be possible to monitor the advancement of the disease.

I believe that with more consistent datas and more information (and possibly more records), it is possible to meliorate the algorithm, and be more precise and accurate in the predictions and monitoring.

5 Sources and references

Initial study

"Improved spiral test using digitized graphics tablet for monitoring Parkinson's disease" by Muhammed Erdem Isenkula, Betul Erdogdu Sakarb, Olcay Kursuna :

https://pdfs.semanticscholar.org/83f6/c11e9ebab1dea4aa3c5a7f9eb692f33d17c1.pdf

Data set, UCI :

https://archive.ics.uci.edu/ml/datasets/Parkinson+Disease+Spiral+Drawings+Using+Digitized+Graphics+Tablet

Brain anatomy :

https://en.wikipedia.org/wiki/Midbrain

Definitions :

https://www.yourdictionary.com

Dopamine :

https://en.wikipedia.org/wiki/Dopamine

Neurotransmitter :

https://en.wikipedia.org/wiki/Neurotransmitter

Parkinson's disease :

https://en.wikipedia.org/wiki/Parkinson%27s_disease

Substantia nigra :

https://en.wikipedia.org/wiki/Substantia_nigra