

# ParkinsonDisease\_SpiralDrawing

*Nina Caparros*

*2019-11-14*

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
I	Nota bene . . . . .	3
II	Parkinson's disease . . . . .	3
III	Project overview . . . . .	4
A	Material . . . . .	5
B	Tests . . . . .	5
C	Archimedean spiral . . . . .	7
D	Goal . . . . .	7
IV	Dataset overview . . . . .	8
<b>2</b>	<b>Method and analysis</b>	<b>8</b>
I	Initial Data . . . . .	8
II	Data cleaning . . . . .	9
A	Format . . . . .	9
B	Cleaning TimeStamp . . . . .	10
C	Calibrating the samples . . . . .	12
III	Data analysis . . . . .	20
A	Global analysis . . . . .	20
B	Archimedean Spiral . . . . .	20
C	Static Spiral Test analysis . . . . .	20
D	Dynamic Spiral Test analysis . . . . .	20
E	Stability Test on Certain Point analysis . . . . .	20
IV	Issues . . . . .	20
V	Prediction algorithms . . . . .	20
A	Creating training and testing sets . . . . .	20
B	Parameters . . . . .	20
<b>3</b>	<b>Results</b>	<b>20</b>
<b>4</b>	<b>Conclusion</b>	<b>20</b>
<b>5</b>	<b>Glossary</b>	<b>20</b>
<b>6</b>	<b>Sources and references</b>	<b>21</b>

# 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 is 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 terms annotated with an asterisk are explained or more detailed in the glossary at the end of the report.

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](https://github.com/ncaparros/ParkinsonDisease_SpiralDrawing)).

## II Parkinson’s disease

Parkinson’s disease, sometimes abbreviated to 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).

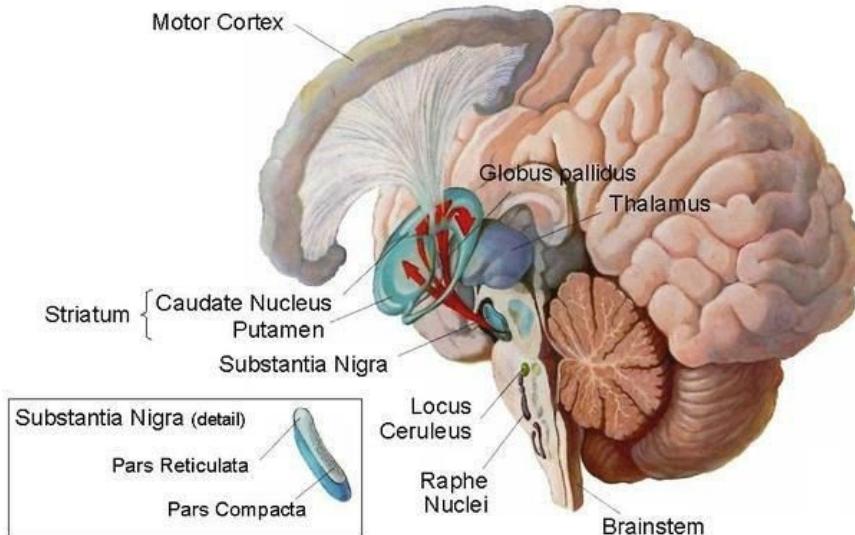
The disease affects mostly the motor system, as tremor, akinesia\*, shaking, rigidity, slowness of movement, difficulty with walking,... and as it worsen it can cause depression, anxiety (more than a third of people with Parkinson’s disease), emotional and sleep troubles, and in the advanced stages the 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 2). The *substantia nigra* is a basal ganglia\* divided into two parts : the *pars reticula*\* and the \**pars compacta*\*\* (see Figure 1). 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 a 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 substantia nigra than on a healthy subject, see Figure 2) provokes emotional troubles as said previously, and since the downsized *substantia nigra*

## Brain Regions Affected by Parkinson's Disease



Parkinson's disease

Figure 1: Lateral cross-section of the brain (source : <http://www.neuroconvention.com/>)

is connected to the motor cortex (via the *pars reticula*, see Figure 1), it causes the parkisonian syndromes.

The Figure 1 shows a lateral cross-section of a brain. The red arrows represent the dopamine's exchanges between the *pars reticula*. The Figure 2 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 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 Erdogan Sakar and Olcay Kursun) which purpose was to monitor Parkinson's disease with digitized graphics tablet. The goal of this

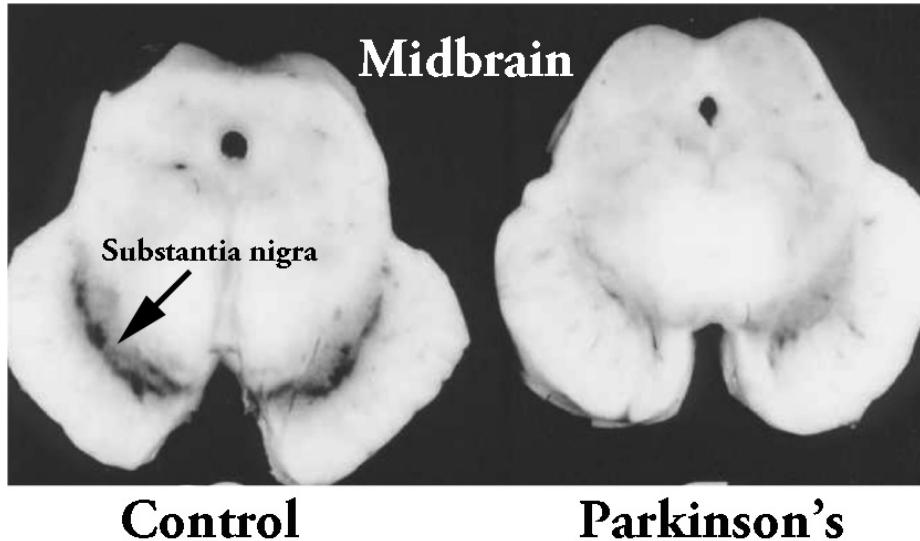


Figure 2: *Substancia nigra* differences between a healthy brain and a Parkinson's brain (source : <https://scienceofparkinsons.com/>)

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 3). The tablet would measure several parameters as : the coordinates (x-y-z) of the pen on the table, the pressure over the screen, the grip angle on the pen at regular time intervals. A software was developed in order to test the coordination of the patient.

## B Tests

The three tests performed by the patients were :

- **Static Spiral Test (SST)** is a traditional test usually performed with paper and pencil. An Archimedean spiral (see following plot) 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)** is a new test introduced in the study,



Figure 3: Wacom Cintiq 12WX graphics, source : <https://www.bhphotovideo.com/>

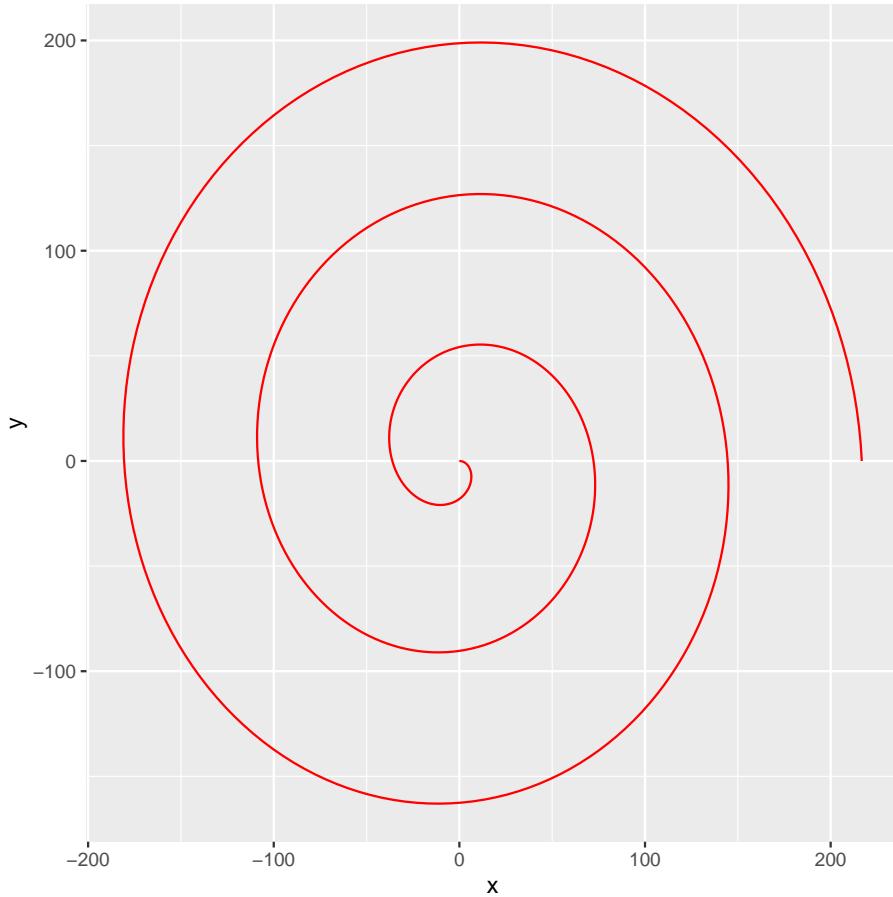


Figure 4: Archimedean spiral

where the archimedean spiral *blinks*. It is only seen at certain times. It becomes more difficult to follow the spiral.

- **Stability Test on Certain Point** is 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 determine the patient's hand stability and hand tremor level.

### C Archimedean spiral

### D Goal

In this report I will try to 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 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.

## 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 :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 described 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 previous section), the first step was to add a random identifier to each patient and to note if he has Parkinson's disease or not. The random identifier had 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 allowed to not get focused on the patient id.

```
##                                     V1 patientID isPwp
## 1 200;204;0;73;910;1732647300;0 3963734496 FALSE
## 2 200;204;0;218;900;1732647307;0 3963734496 FALSE
```

```

## 3 200;204;0;253;900;1732647314;0 3963734496 FALSE
## 4 200;204;0;304;900;1732647321;0 3963734496 FALSE
## 5 200;204;0;351;900;1732647328;0 3963734496 FALSE
## 6 200;204;0;386;900;1732647335;0 3963734496 FALSE

```

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

```

##   X   Y Z Pressure GripAngle   Timestamp TestID patientID isPwp
## 1 200 204 0      73        910 1732647300      0 3963734496 FALSE
## 2 200 204 0     218        900 1732647307      0 3963734496 FALSE
## 3 200 204 0     253        900 1732647314      0 3963734496 FALSE
## 4 200 204 0     304        900 1732647321      0 3963734496 FALSE
## 5 200 204 0     351        900 1732647328      0 3963734496 FALSE
## 6 200 204 0     386        900 1732647335      0 3963734496 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 was then created and filled as :

```

##   patientID isPwp
## 1 3963734496 FALSE
## 2 792736254 FALSE
## 3 1714771463 FALSE
## 4 1180118019 FALSE
## 5 2105782496 FALSE
## 6 3001810178 FALSE

```

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

## II Data cleaning

In this subsection I explained 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.

```

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 were all converted as numeric.

```

df <- df %>% mutate(X = as.numeric(X),
                      Y = as.numeric(Y),
                      Z = as.numeric(Z),
                      Pressure = as.numeric(Pressure),
                      GripAngle = as.numeric(GripAngle),
                      Timestamp = as.numeric(Timestamp))

```

## B Cleaning TimeStamp

The timestamp columns seemed pretty obscure, 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`,...).

```

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 subtracted the first timestamp of every couple test/patient to all the timestamp values, making the first value 0. To do this I had to create a new empty data frame, and two for loops : one for the test (0 to 2) and one for the patients (0 to `nrow(patients)`). 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.

```

head(df)

##      X   Y   Z Pressure GripAngle   Timestamp TestID patientID isPwp
## 1 200 204  0       73      910 1732647300      0 3963734496 FALSE
## 2 200 204  0      218      900 1732647307      0 3963734496 FALSE
## 3 200 204  0      253      900 1732647314      0 3963734496 FALSE
## 4 200 204  0      304      900 1732647321      0 3963734496 FALSE
## 5 200 204  0      351      900 1732647328      0 3963734496 FALSE
## 6 200 204  0      386      900 1732647335      0 3963734496 FALSE

completeDf <- data.frame()

#For each of the tests "test"
for(test in seq(0, by=1, length=3)){

  #For each of the patients "indPatient"
  for(indPatient in seq(1, by=1, length=nrow(patients))){

    #Create a temporary data frame for patient "indPatient"
    #and test "test"
    temp_df <- df %>%
      filter(TestID == test &
             patientID == patients[indPatient,]$patientID) %>%
      arrange(Timestamp)

    #Get first value of timestamp
    initialTimestamp = temp_df[1,]$Timestamp

    #Mutate Timestamp so that the very first value of Timestamp
    #for patient "indPatient" and test "test" is equal to 0
    temp_df <- temp_df %>%
  }
}
```

```

    mutate(Timestamp = Timestamp - initialTimestamp)

#Bind temporary data frame to complete data frame
completeDf <- rbind(completeDf, temp_df)

}

}

head(completeDf)

##      X   Y Z Pressure GripAngle Timestamp TestID patientID isPwp
## 1 200 204 0       73      910        0      0 3963734496 FALSE
## 2 200 204 0      218      900        7      0 3963734496 FALSE
## 3 200 204 0      253      900       14      0 3963734496 FALSE
## 4 200 204 0      304      900       21      0 3963734496 FALSE
## 5 200 204 0      351      900       28      0 3963734496 FALSE
## 6 200 204 0      386      900       35      0 3963734496 FALSE

```

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

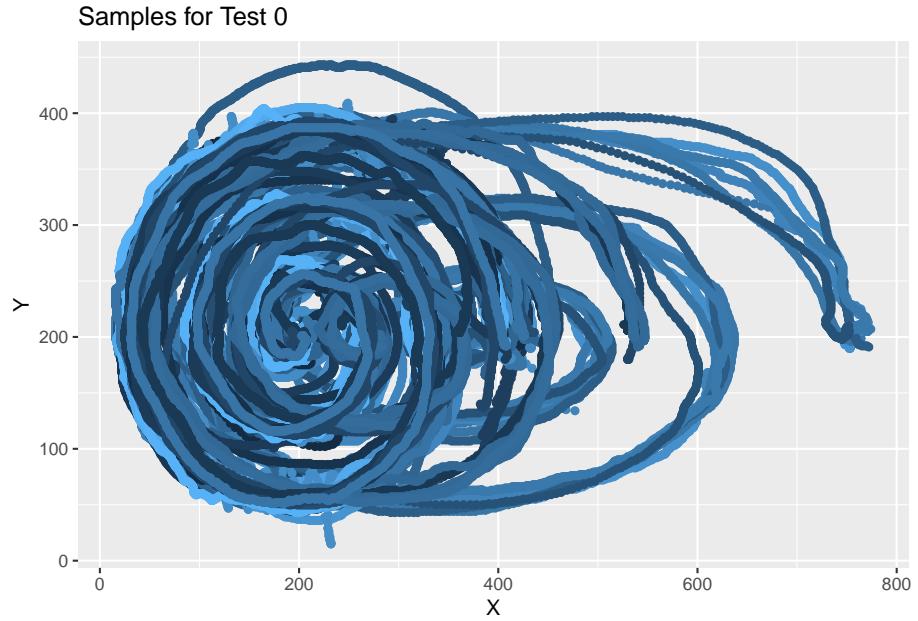
## C Calibrating the samples

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

```

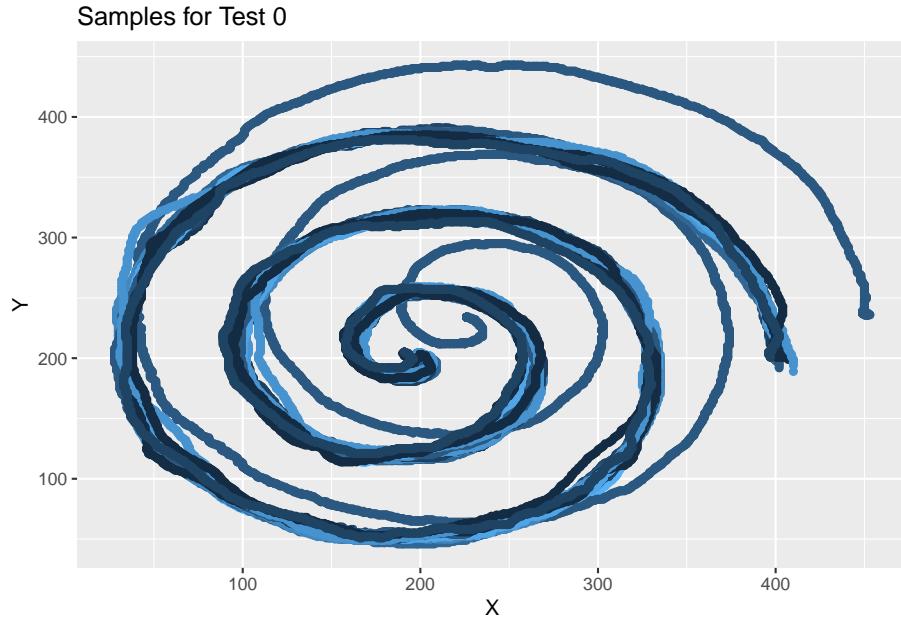
completeDf %>%
  filter(TestID == 0) %>%
  ggplot() +
  geom_point(aes(X,Y, color=patientID),
             show.legend = FALSE) +
  ggtitle("Samples for Test 0")

```



Another odd detail was a spacial gap between patterns with the same ratio (see Plot). One of the control sample seemed to have been shifted.

```
completeDf %>%
  filter(TestID == 0 &
         isPwp == FALSE) %>%
  ggplot() +
  geom_point(aes(X,Y,
                 color=patientID),
             show.legend = FALSE) +
  ggtitle("Samples for Test 0")
```

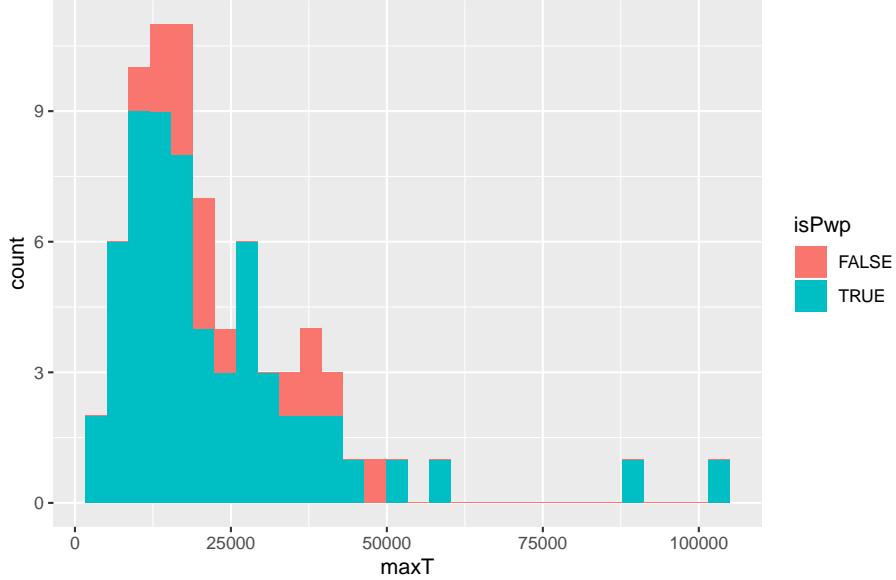


This disparity between the samples prevented me from trying to analyse the spacial differences between the generated Archimedean spiral and the patient's drawing. I checked if the timestamp data seemed affected or not (see Plot). It seemed as only the X and Y values were sometimes off.

```
completeDf %>%
  filter(TestID == 0) %>%
  group_by(patientID) %>%
  summarize(maxT = max(Timestamp),
           isPwp = first(isPwp)) %>%
  ggplot() +
  geom_histogram(aes(maxT,
                     fill=isPwp),
                 stat="bin")+
  ggtitle("Sample Timestamps for Test 0")

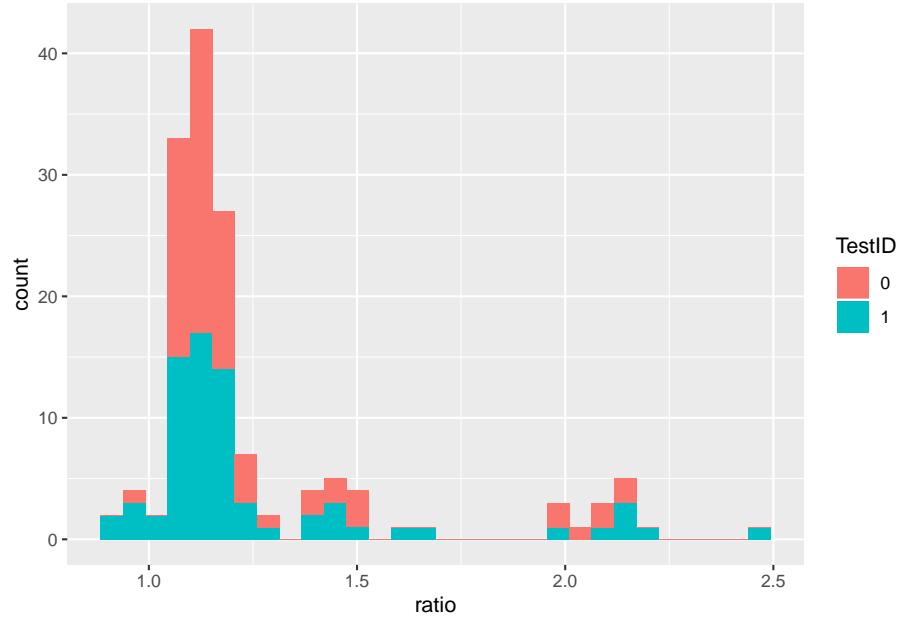
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Sample Timestamps for Test 0



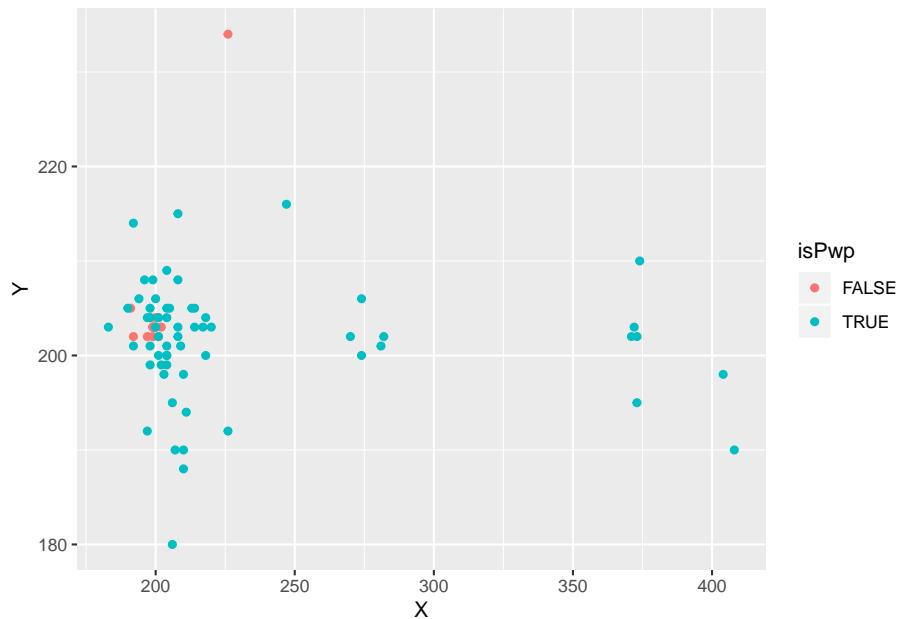
To calibrate the X and Y values, the ratios  $\frac{X}{Y}$  were calculated and stored in a new data frame  $ratiosDf$ , by patient and test. This data frame contained the ratio between X and Y as  $ratio = \frac{X_{max} - X_{min}}{Y_{max} - Y_{min}}$

```
ratiosDf <- df %>%
  group_by(patientID, TestID) %>%
  summarize(ratio=(max(X)-min(X))/(max(Y)-min(Y)),
            distY= max(Y)-min(Y),
            distX= max(X)-min(X),
            isPwp = first(isPwp))
ratiosDf %>% filter(TestID != 2) %>%
  ggplot() +
  geom_bar(aes(ratio, fill=TestID), stat="bin")
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We can indeed see that even if most ratios are around 1.3, some outliers are around 2. Even the initial point of the drawing was different (see next plot).

```
completeDf %>% filter(TestID == 0 & Timestamp == 0) %>% ggplot() + geom_point(aes(X,Y, color = isPwp))
```



To standardize all the spirals, several parameters were taken into account : \*

The ratio  $(\frac{X_{max} - X_{min}}{Y_{max} - Y_{min}})$  should be of one \* 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  $X_{standardized_i} = \frac{X_{old_i}}{ratio}$  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 :  
 $X_{resized_i} = \frac{X_{old_i}}{SpiralLength} * Size$  and  $Y_{resized_i} = \frac{Y_{old_i}}{SpiralLength} * Size$  when  
 $SpiralLength = SpiralHeight$  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 assumed it was  $400px$ . 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 :

- $X_{param} = X_{mean}$  and  $Y_{param} = Y_{mean}$
- $X_{param} = X_{median}$  and  $Y_{param} = Y_{median}$
- $X_{param} = mean(X_{Y=Y_{median}})$  and  $Y_{param} = mean(Y_{X=X_{median}})$
- $X_{param} = mean(X_{Y=Y_{mean}})$  and  $Y_{param} = mean(Y_{X=X_{mean}})$
- $X_{param} = mean(X_{Y=Y_{Timestamp=0}})$  and  $Y_{param} = mean(Y_{X=X_{Timestamp=0}})$

```
completeCalibratedDf <- data.frame()

for(test in seq(0, by=1, length=3)){
  for(indPatient in seq(1, by=1, length=nrow(patients))){
    ratioLines <- ratiosDf %>%
      filter(patientID == patients[indPatient,]$patientID &
             TestID == test)

    if(nrow(ratioLines) >=1){

      ratio <- ratioLines[1,]$ratio
      distY <- ratioLines[1,]$distY

      temp_df <- completeDf %>%
        filter(TestID == test &
               patientID==patients[indPatient,]$patientID) %>%
        arrange(Timestamp)

      pX <- (temp_df[which.min(temp_df$Timestamp),]$X + min(temp_df$X) + max(temp_df$X))/3

      pY <- (temp_df[which.min(temp_df$Timestamp),]$Y+ min(temp_df$Y) + max(temp_df$Y))/3
    }
  }
}
```

```

temp_df <- temp_df %>%
  mutate(X = ((X - pX)/ratio)/distY*400,
        Y = (Y - pY)/distY*400)

completeCalibratedDf <- rbind(completeCalibratedDf, temp_df)
}

}

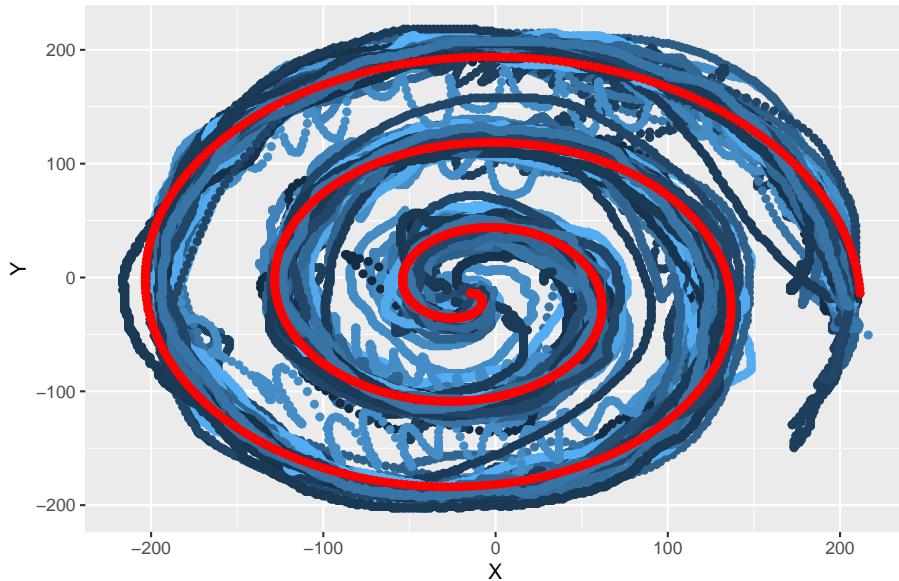
Xadjust <- completeCalibratedDf %>%
  filter(TestID != 2 &
         Timestamp==0 &
         isPwp == FALSE) %>%
  pull(X) %>%
  mean()
Yadjust <- completeCalibratedDf %>% filter(TestID != 2 & Timestamp==0 & isPwp == FALSE) %>%

b <- 12
t <- seq(0,6*pi, length.out=1100)
x <- b*t * (cos(t)) + Xadjust
y <- b*t * -(sin(t)) + Yadjust
archimedeanSpiral = data.frame(x,y)

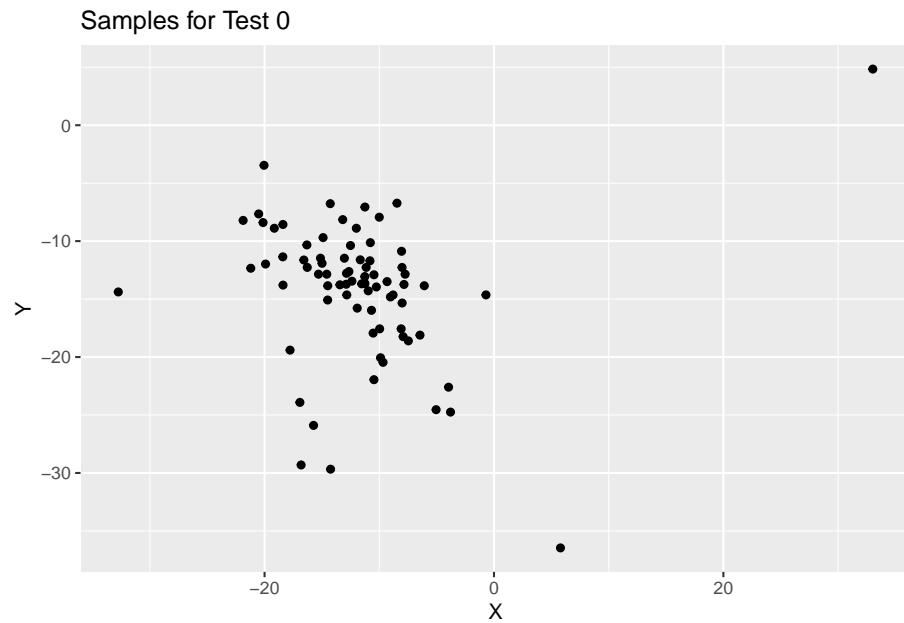
completeCalibratedDf %>%
  filter(TestID == 0) %>%
  ggplot() +
  geom_point(aes(X,Y,
                 color=patientID),
             show.legend = FALSE) +
  ggtitle("Samples for Test 0") +
  geom_point(data=archimedeanSpiral, aes(x,y), color="red")

```

Samples for Test 0



```
completeCalibratedDf %>%
  filter(TestID == 0 & Timestamp==0) %>%
  ggplot() +
  geom_point(aes(X,Y),
             show.legend = FALSE) +
  ggtitle("Samples for Test 0")
```



### **III Data analysis**

- A Global analysis**
- B Archimedean Spiral**
- C Static Spiral Test analysis**
- D Dynamic Spiral Test analysis**
- E Stability Test on Certain Point analysis**

### **IV Issues**

### **V Prediction algorithms**

- A Creating training and testing sets**
- B Parameters**

## **3 Results**

## **4 Conclusion**

## **5 Glossary**

**Akinesia :**

**Basal ganglia :**

**Dopaminergic :**

**Pars compacta :**

**Pars reticula :**

**Substancia nigra :** # Table of figures

## 6 Sources and references

Brain anatomy : <https://en.wikipedia.org/wiki/Midbrain> 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](https://en.wikipedia.org/wiki/Parkinson%27s_disease) Substantia nigra : [https://en.wikipedia.org/wiki/Substantia\\_nigra](https://en.wikipedia.org/wiki/Substantia_nigra)