11-19-2018

Statistical Multi-way Relationships of Human Brain Activity

Chen Feng
cf8805@rit.edu

R·I·T

Statistical Multi-way Relationships of Human Brain Activity

by

Chen Feng

A Thesis Submitted in Partial Fulfillment of the Requirements for

the Degree of Master of Science in Applied Statistics

School of Mathematical Sciences, College of Science

Rochester Institute of Technology

Rochester, NY

November 19, 2018

Committee Approval:

Dr. Peter Bajorski Date

School of Mathematical Science

Thesis Advisor

Dr. Andrew Michael Date

Duke Institute for Brain Science

Committee Member

Dr. Nathan Cahill Date

School of Mathematical Science

Committee Member

ABSTRACT

The human brain activity is a popular and important topic in the medical science field and academic studies. In recent years, scientists have been applying various statistical methods to analyze human brain activity. Correlation between brain regions is the most common and fundamental method used to perform this task. However, correlation describes only a two-way relationship. This work explores a new approach by analyzing multi-way relationships. Due to computational complexities, we concentrate on three-way relationships. In particular, we compare conventional two-way correlations and three-way regression models. Data transformed and processed from 3,280 MRI scans of the human brain are used in modeling and analysis. The results of this research show qualified three-way relationships which have a significant advantage relative to their corresponding two-way relationships. The algorithm proposed in this paper can potentially outperform the conventional two-way correlations in exploring the activity of human brain regions.

CONTENTS

1 INTRODUCTION

1.1 INSPIRATION

How does the human brain work? This is a question for almost every scholar field. From philosophers to doctors, economists to scientists, even for the general public, it is one of the ultimate questions which humans are eager to understand.

The activity of a human brain can be detected through magnetic resonance imaging (MRI) scans. Based on the knowledge of anatomy and imaging processing, scientists can obtain frequency data which represent the activity of different brain regions. With these frequency data, statistical and mathematical researchers can conduct further analysis.

There is plenty of great scientific research regarding the relationship between two different brain regions. Each study has its own creative and unique approaches. In the field of statistics, the most mature method is based on “correlation”. Through this one number, it is possible to judge whether the activity of two brain regions is mutually promoted or restrained. Therefore, scientists have obtained some insights on how our brain is internally related. However, it is reasonable to think that there exist more complex or multi-dimensional internal relationships between human brain regions.

1.2 GOALS OF WORK

In this thesis, a new way is proposed to explore the multi-way relationships, which involves multiple brain regions in one statistical model, instead of just two regions, as is done in the traditional approaches using correlations. Here we will concentrate on three-way relationships, but the same approach can be used for four and more-way relationships in the future work.

The modeling chooses one brain region as the dependent variable, and two other brain regions as independent variables. By measuring how much advantage the three-way model has, compared to the corresponding two-way models (correlations), this measurement named “D value” calculated from subtracting maximum squared correlation of the two-way models from the R-square of the three-way model. Test on all possible 253,460 variable selection combinations, and recursively modeling on each scan by small pieces of frequency data. It is feasible to find out for each model, how constantly the D-values are high on one scan. This measurement is named “consistency”, and it is derived from calculating the percentage of the length of time points with high D-value over the total length of a scan.

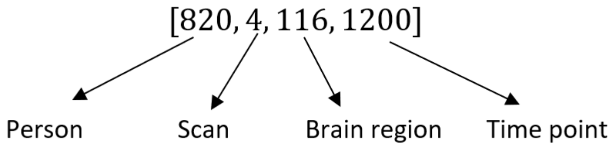
Except for horizontal modeling and calculation on each scan and each model, this thesis also provides variance analysis on consistency which is conducted vertically between scans and modeling on choosing “person” as a random factor.

These methods provide an effective path to understand and evaluate the models as well as a new statistical perspective to analyze human brain activity.

1.3 DATA STRUCTURE AND VARIABLE COMBINATION

The original experiment contains MRI scans from 820 people, each person was scanned four times. Then, the MRI images were converted into frequency data through complex imaging process. During this process, the original MRI images were separated and recognized as 116 brain regions. Each brain region contains 1,200 points of the frequency data.

The data is stored into an array, whose dimensions are shown as follow:



To start the three-way modeling, we build a model based on the data from the first scan of the first person.

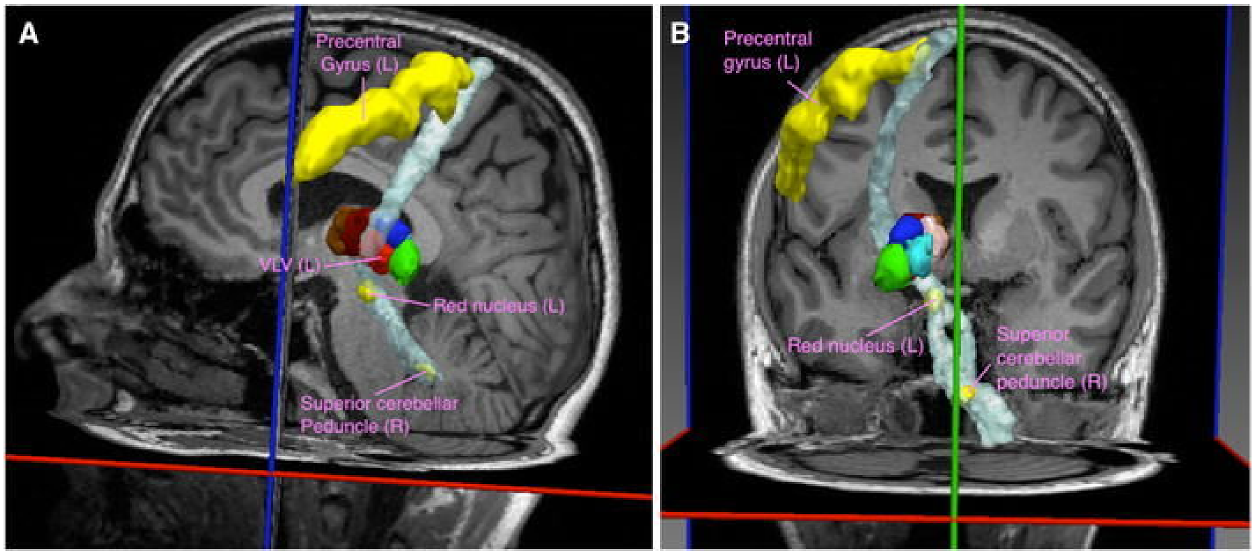
$$First\ person\ first\ scan:$$

$$[1, 1, 1{:}\,116, 1{:}\,1200]$$

For exploring purpose, fix the dependent variable (Z) as left precentral gyrus (region 1),

$$Z: [1, 1, 1, 1{:}\,1200]$$

shown as the yellow part in the following graph:



Graph 1.3

Battistella, G., Najdenovska, E., Maeder, P. et al. Brain Struct Funct (2017) 222: 2203. [1]

The dependent variables (X and Y) will be randomly selected from the other brain regions except region 1. Mathematically, the total number of possible three-way models should be: $C^2_{(116-1)}$, which equals 6,555.

2 THE ADVANTAGE OF THREE-WAY RELATIONSHIPS

There is a two-layer system to decide the quality of a model, the first layer is “D-value” and the second layer is “Consistency”. Consistency is calculated based on D-value.

2.1 D-VALUE

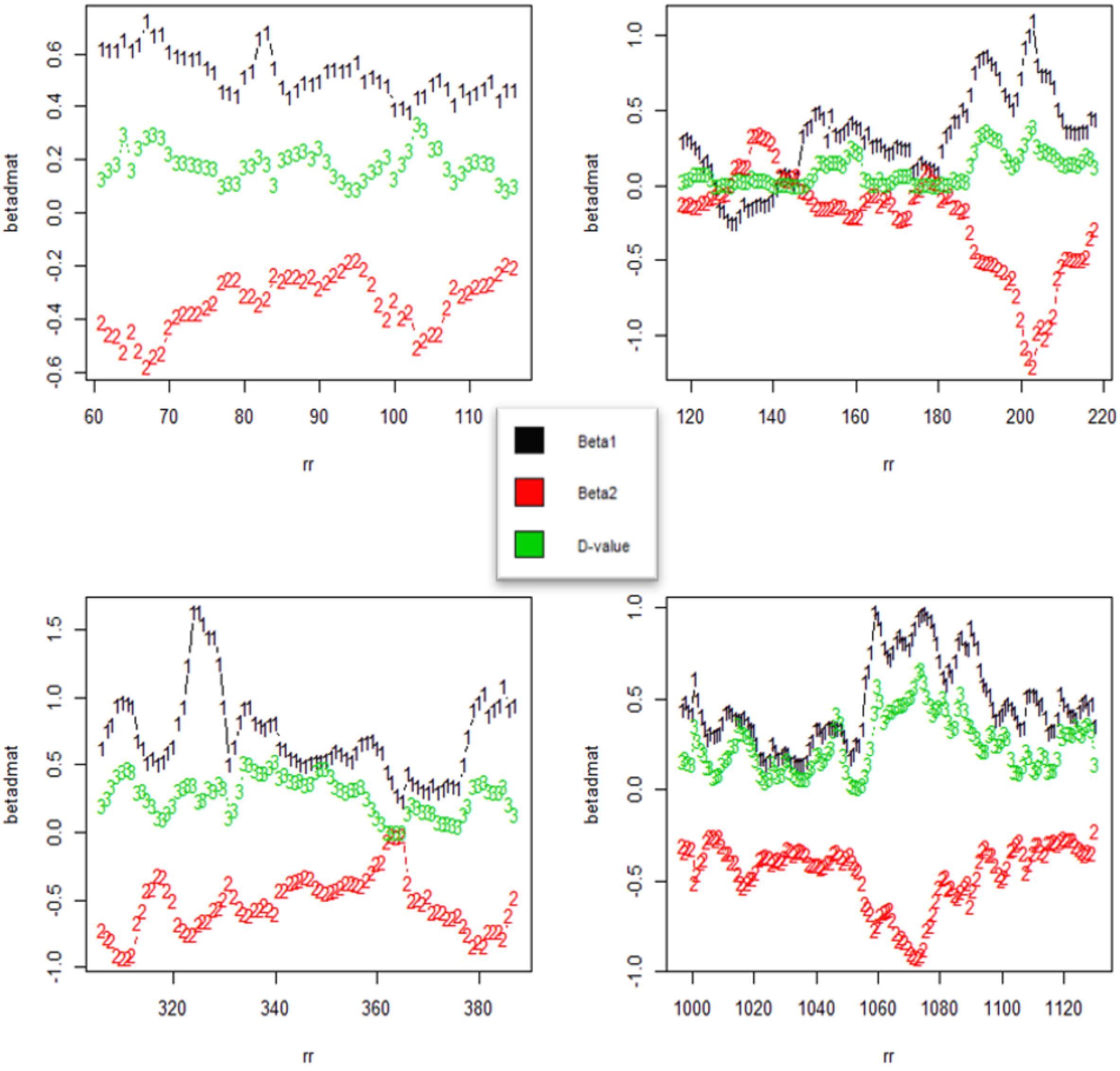
The D-value is the most basic and core concept for all the modeling process involved in this thesis. D-value is calculated by subtracting the maximum squared correlation of the two-way models from the R square of the three-way model. It represents the relative advantage of a three-way model over the corresponding two-way models. Since in the modeling process and comparison, we are basically adding an extra independent variable to a two-way model to form a three-way model, the three-way model must be more “accurate” than the two-way model, thus, R square will always be larger than the maximum squared correlation, so that the D-value will always be positive.

Considering the meaning of D-value, the larger its value, the more advantage the three-way model has over the corresponding two-way models.

2.2 COEFFICIENTS AND D-VALUE

During the process of exploring the potential relationship between D-value level and other critical parameters, like coefficients of the two independent variables in the three-way model, p-value, and constant item, we find out that where D-value appears to stay at a high level, the coefficients values are opposite in sign. Usually, the coefficient of the first independent variable is negative, and the coefficient of the second independent variable is positive, then the D-value at that point is high.

Coefficients and D-value Plots by Range



Graph 2.2 Zoom in range plot of coefficients and D-value

2.3 CONSISTENCY

In addition to discovering the occurrence of extremely high D-value at each single time point, it is crucial to measure how often the D-value stays above an acceptable and relatively high level as well. Because the goal of this thesis and work is not to find models that provide good fit at only a few time points, but to find models that provide stable and accurate performance.

Consistency is essentially a ratio that represents what is the portion of time points that return acceptable D-values vs. the total number of time points on a single scan. High consistency shows a model has stable performance for this scan.

The conditions of consistency contain three relevant parameters: D-value, the coefficient of the first independent variable and the coefficient of the second independent variable. Also, all of the parameters involved have their own thresholds. The reason to choose these parameters and the reason to set the thresholds to a certain level will be explained in detail when the model selection is done in Section 4.

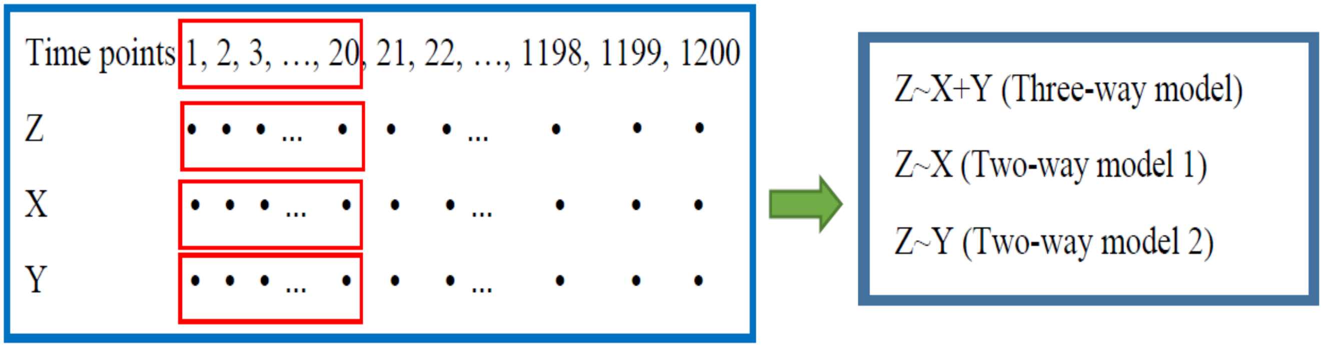
2.4 STRENGTH

Consistency measures the performance of a multi-way model over a scan based on an acceptable level of threshold. Strength, however, by setting a relatively higher threshold, is focused on measuring how frequently a multi-way model can produce high D-value over a scan. Strength helps us to have a more comprehensive understanding of the level and distribution of D-value. Since the structure and parameters for strength function are more restrictive but are set to be similar to consistency, the strength is mainly used to support and assist with the post-stage evaluation when qualified models are selected out.

3 THREE-WAY RELATIONSHIP MODELING

3.1 WINDOW SIZE AND RECURSIVE FITTING

Because the data have time series characteristic, to better represent the change of critical measurements over time, instead of fitting a model with the entire data from one scan, a window whose size equals 20 is selected to assist with the modeling process. On each scan, there are multiple models built, and this is a recursive process:



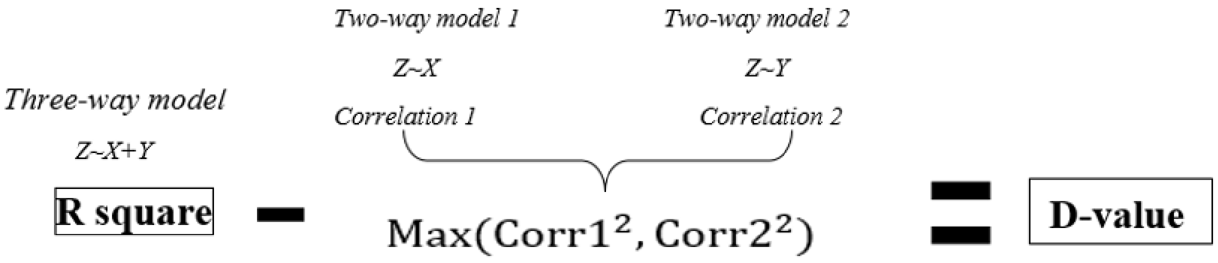
Graph 3.1 Recursive fitting process

Move the red window to the right, one position at a time, build a linear three-way model and two-way models with the data in the window after each move. After finishing this recursive modeling process, there are 1,181 groups of three-way fits and corresponding two-way fits. But we only include 1180 groups of fits, the last group was excluded from further calculations in order to avoid the edging effect.

As explained at the end of section 1.3, with a fixed dependent variable, the total number of three-way models is 6,555, and for each model, on the first scan, there are 1,180 group of fits, each group of fits has one D-value.

3.2 D-VALUE CALCULATION

As what was mentioned in the previous part, D-value is calculated by subtracting the maximum squared correlation of the two corresponding two-way models from the R square of the three-way model. See the following graph for a better understanding of the D-value calculation process:



Graph 3.2 Calculate D-value

The D-values are stored into a 1,180 by 6,555 matrix, each column represents the D-values from one model. Having a concept of how the calculated results are stored helps readers to obtain a better understanding of the upcoming analysis.

3.3 CONSISTENCY CALCULATION

The process of consistency calculation contains two steps, the first step is based on logical judgments and counting, the second step is division. For the first step, we need an “acceptable threshold for D” (th_a), two other thresholds for the coefficients beta1 and beta2 of the independent variables in the three-way model (th_{b1} and th_{b2}). For each model, it has one consistency on every scan where it is applied on. The following equation shows how to calculate consistency column-wise in the D matrix, where $length(D) = 1180$. This length is related with the window size.

$$Consistency = \frac{length(D > th_a \ \&Beta1 > th_{b1} \ \&Beta2 < th_{b2})}{length(D)} \qquad (3.3.1)$$

The consistency results are simply stored as a vector.

3.4 AN EXAMPLE OF CALCULATIONS IN R

To further demonstrate the actual practice of how to calculate D value and consistency, here is a step by step example with R code.

First of all, the data is stored in a big array, and we need to extract data for one specific scan. To achieve this, two parameters p and s are defined. P represents the person who provides the scan, and for each person, there are four scans, so s is used to indicate which specific scan among the four we are going to extract. Then assign the extracted data to a data frame named “scan” for later use. The scan data is a 1,200 by 116 array. 1,200 is the total moments in one scan and 116 is the total number of brain regions.

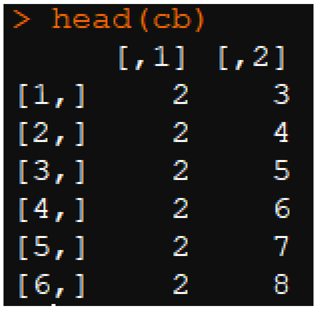
```
p = 1

s = 1

scan <- Scans.arr[ , ,s,p]
```

Now we have the data for one scan extracted and assigned to the data frame, the next step is to build a matrix to store D value. Here, the dependent variable Z is fixed to be brain region 1, for independent variables X and Y, there are $C_{115}^{2} = 6{,}555$ models (k) we will fit. This number will be the vertical dimension, the row number of the matrix. As for the horizontal dimension, since we choose window size (m) as 20, for each model there will be 1,180 fits (m), this is the column number of the matrix.

```
h=20  # window width

k=(115*114)/2  # number of models/number of rows of D matrix

m=1200-h  # number of fits of each model / number of columns of D matrix

dmat <- matrix(data=NA, nrow = k, ncol=m)
```

Then build a data frame for independent variables of different models, this data frame named “cb” and is build and looks like the following:

```
cb <- t(combn(c(2:116),2))
```



```
> head(cb)
     [,1] [,2]
[1,]    2    3
[2,]    2    4
[3,]    2    5
[4,]    2    6
[5,]    2    7
[6,]    2    8
```

With all parameters set up, the final step is performing the recursive calculation by applying two layers nested for loop, the R code is shown below:

```
for (l in 1:k){
  for (i in 1:m){
    z <- scan[i:(i+19), 1]
    x <- scan[i:(i+19), cb[l,1]]
    y <- scan[i:(i+19), cb[l,2]]
    dmat[l,i] <- summary(lm(z~x+y))$r.squared-max((cor(z,x))^2,(cor(z,y))^2)
  }
}
```

Pointers l and i are responsible to trace data piece by moment and by brain region index. After the program finished running, store the dmat into a Rdata file. Thus, we have a D value matrix for all models with fixed Z and designated window size on one specific scan.

4 MODEL SELECTION

4.1 THRESHOLDS FOR D-VALUE

To recognize the qualified models, one feasible method is setting a target D-value threshold (“th”), then count how many D-values along one scan for a specific model are exceeded the target threshold. It was mentioned above that D-value measures the advantage of the three-way model when compared to the best corresponding two-way models. The more the D-values exceeded the target, the better the three-way model’s performance.

We can start from looking for a few good models and research on their characteristics, then use these characteristics to measure other models.

In order to filter out “high-quality three-way model”, a high threshold is set for filter purpose:

$$\text{th} = 0.5$$

The next step is to count, for each model, how many D-value on the first scan exceeded the high threshold we set.

After comparison and calculation, model 4413 returns the greatest number of D-value which exceeded the high threshold, the model (#4413) contains the following variables:

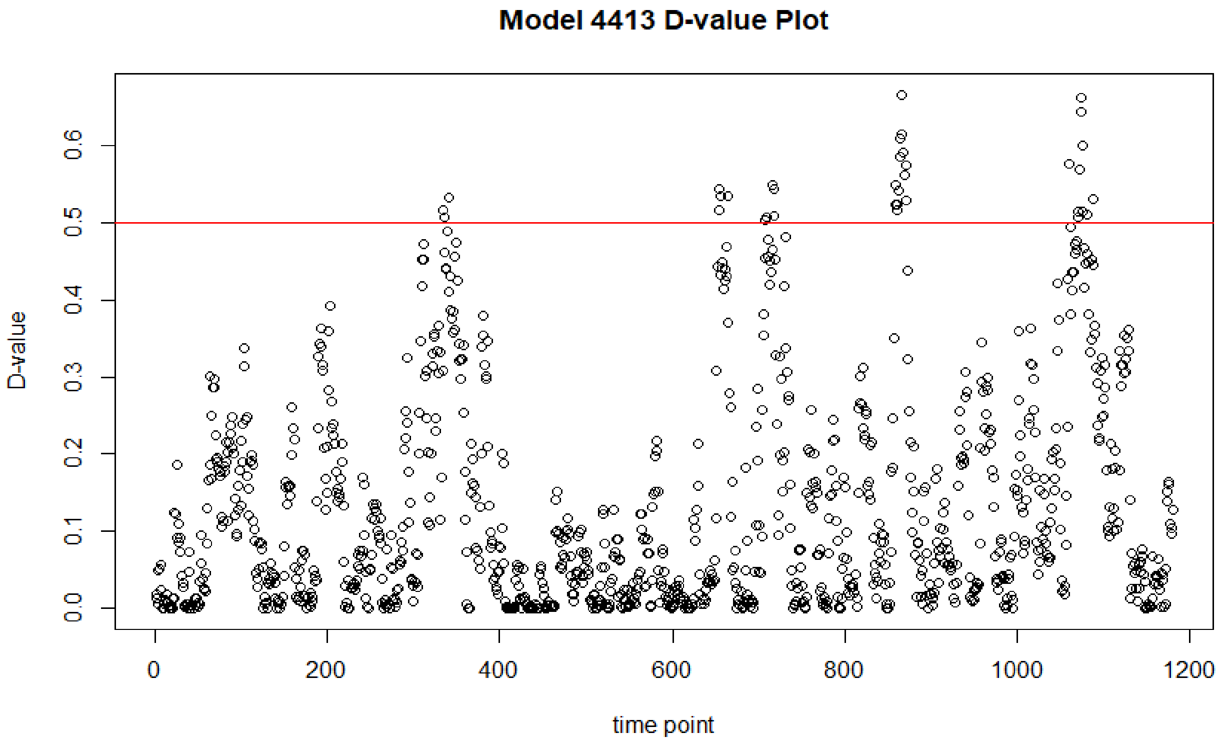
Dependent variable (Z): left precentral gyrus

Independent variable (X): Left middle occipital gyrus

(Y): Right Inferior occipital gyrus

It returns 36-time points on the first scan where D-value is greater than the target threshold, which is about 3.05% over the 1,180 D-values. This is the highest level among all models. Now, we start from this model and try to summarize some useful characteristics for further model D-value performance evaluation.

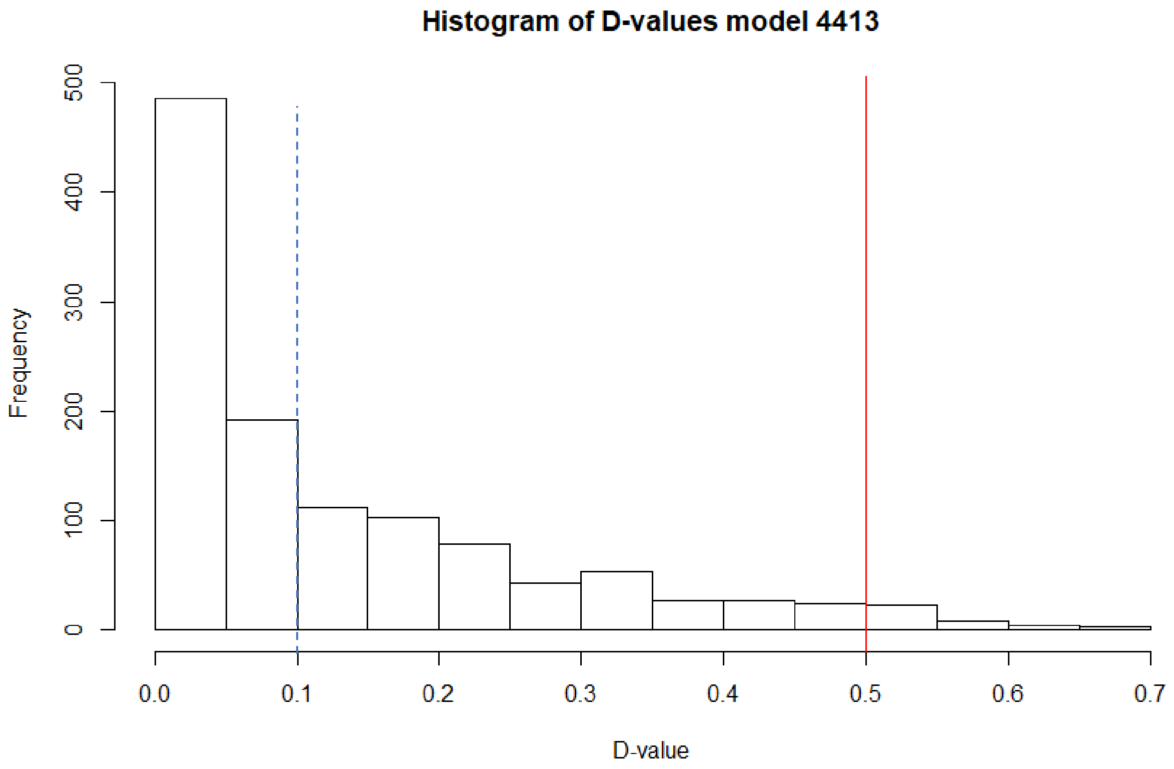
From the following plot, it is clear to see that, D-value exceeds the current target threshold (red horizontal line) occasionally. But the movement of D-value seems to have some "wave" motion, this is because the original frequency data has time serials feature. It also suggests there should exist some factors which affect the three-way model advantage over two-way models.



Graph 4.1.1 D-value of model 4413 for the first scan with threshold 0.5 mark

Model 4413 is one of "extremely good cases" which shows the best three-way model performance.

Then we move to the histogram of the D-values of this model:

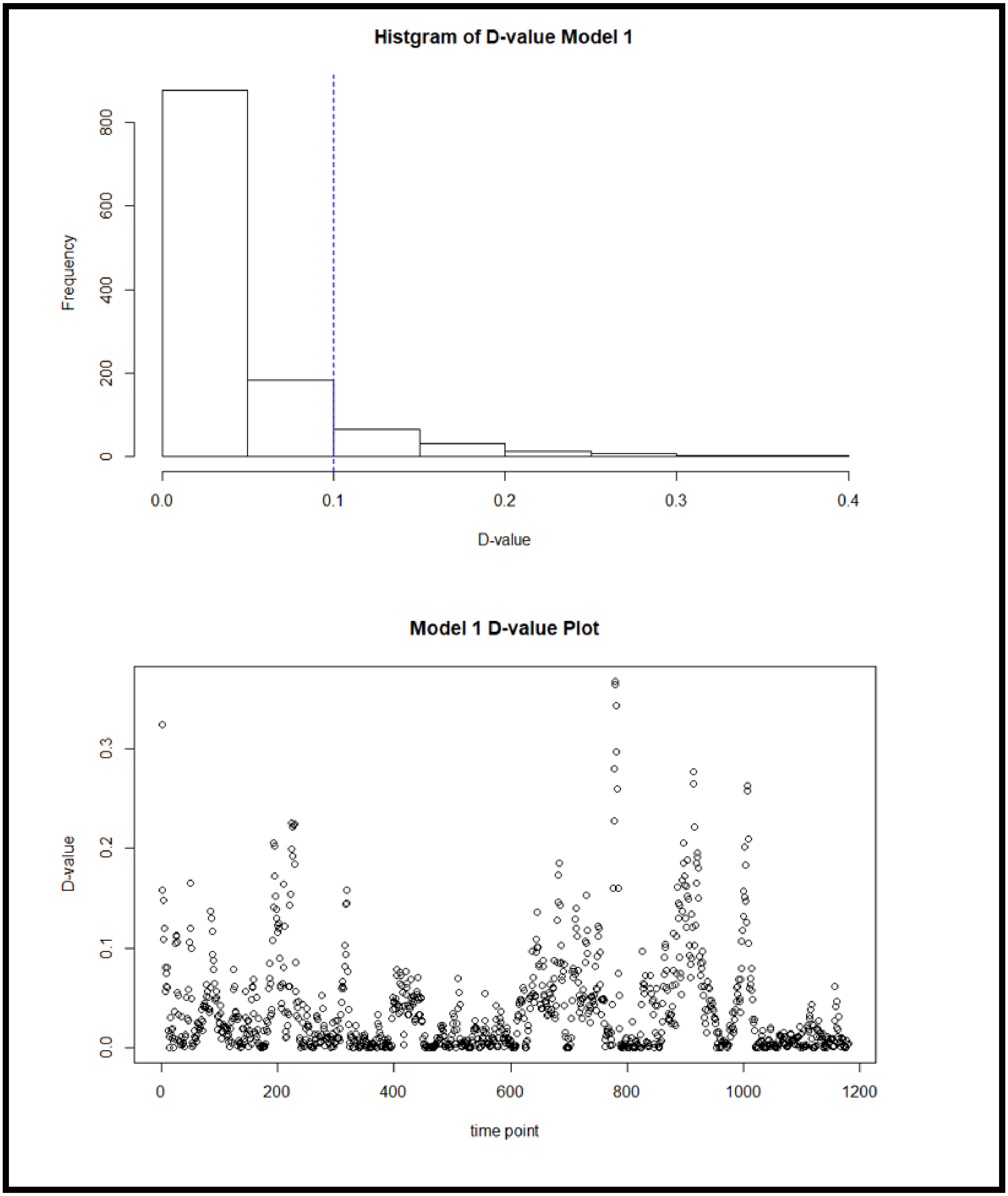


Graph 4.1.2 D-value histogram of model 4413 with marks

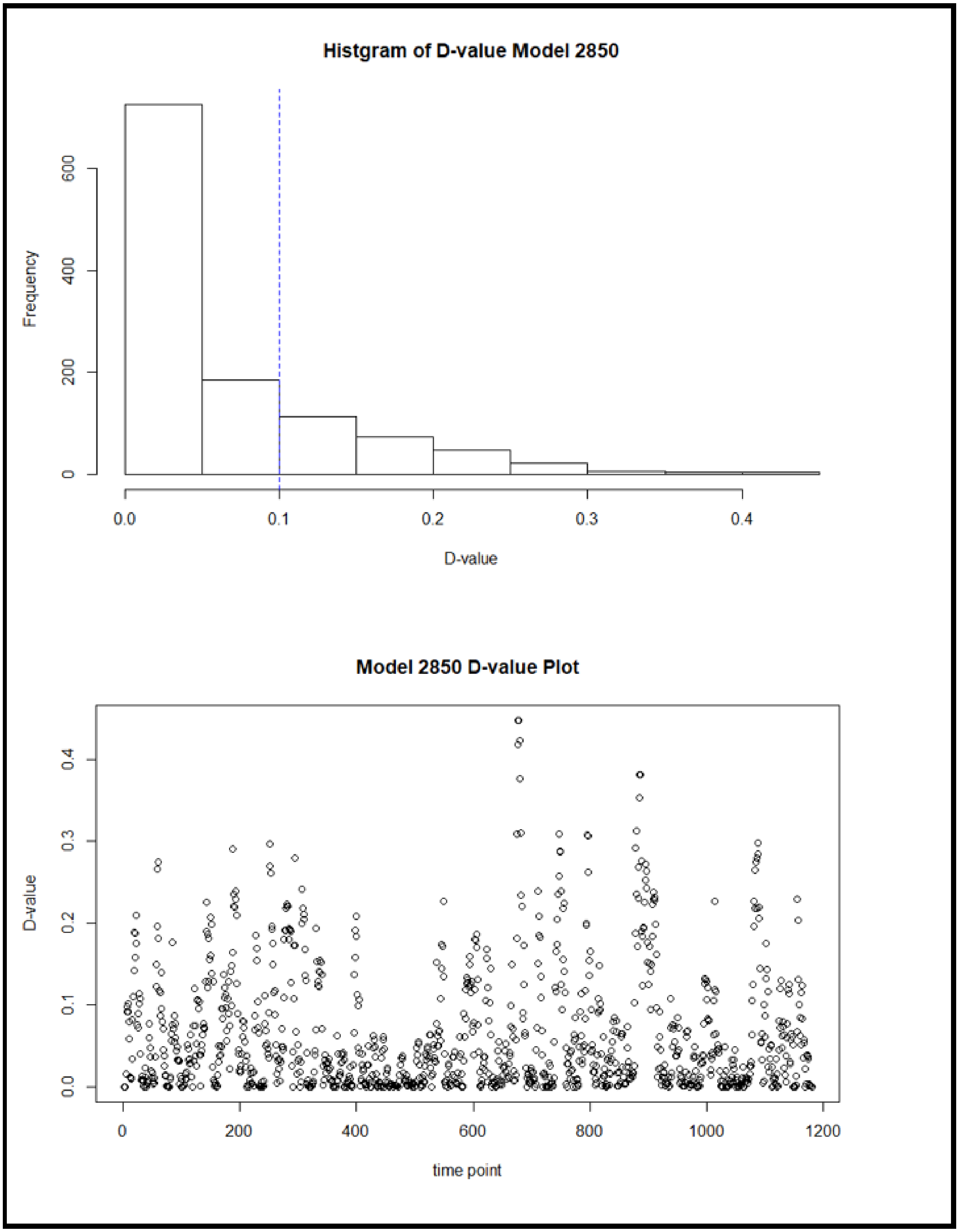
It is obvious to notice that in the D value distribution, most of the D values are below 0.1 (blue vertical dash line), and only a few of them are exceeded 0.5 (red vertical line).

Since 0.5 is a high-level threshold, 0.1 may be a good choice for a more generalized relatively lower level threshold, which will be able to filter out most of the ordinary D-values and include the good D-values at a not too special or too rare level.

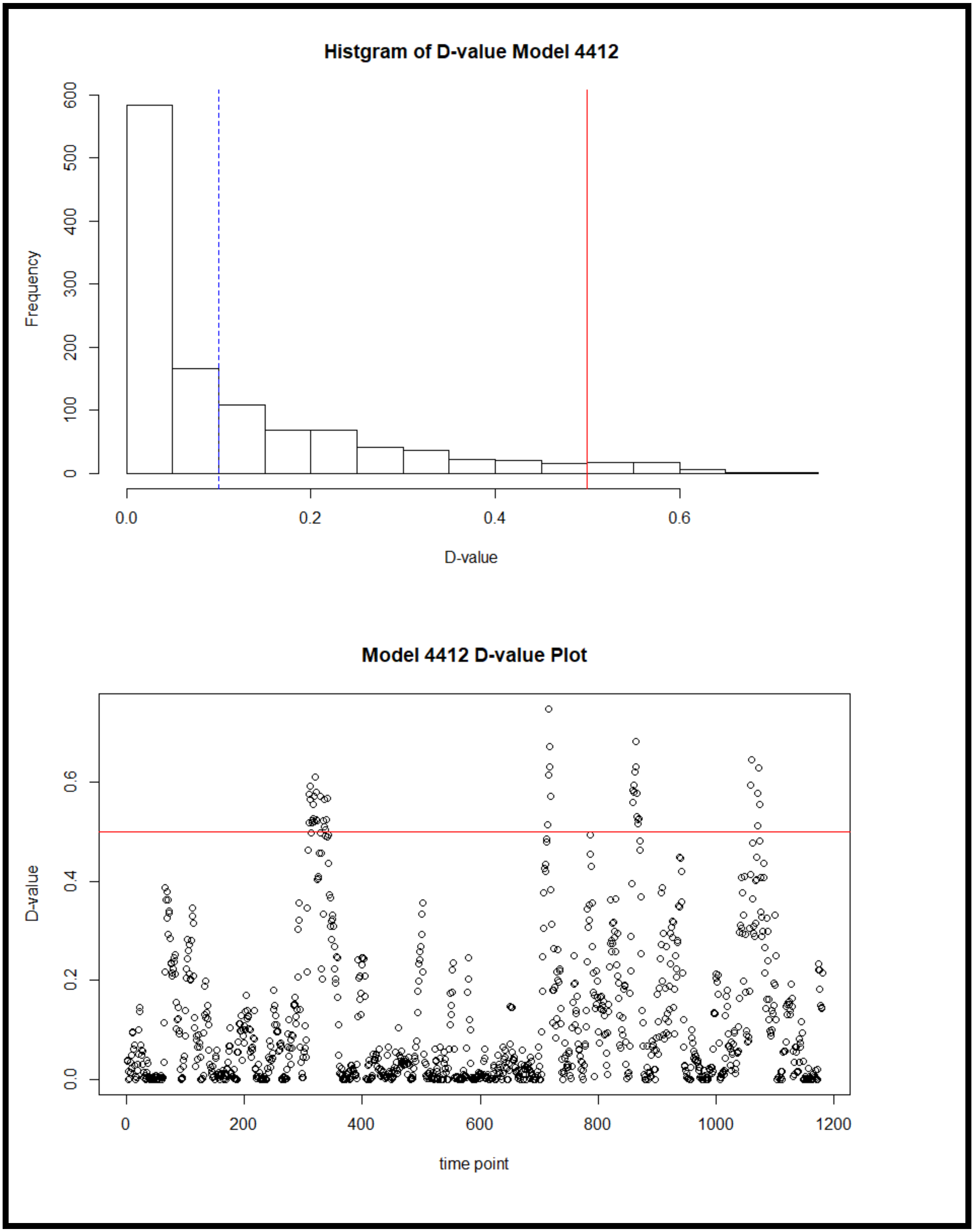
To verify the characteristics on other models, here are several randomly selected models, check the following plots for three randomly selected models, which are model 1, model 2850, and model 4412:



Graph 4.1.3 – 4.1.4 Histogram and Plot of D-value for model 1



Graph 4.1.5 - 4.1.6 Histogram and Plot of D-value for model 2850



Graph 4.1.7- 4.1.8 Histogram and Plot of D-value for model 4412

All of the above three randomly selected models have most D-values below 0.1, one of them has D-values occasionally across 0.5, but other models do not have D-values across 0.5. This is also true for many other models.

The following code, table, and plot show that on the first scan for all of the 6,555 models, the percentage of each model's D-value falls into the range of (0.1, 0.5).

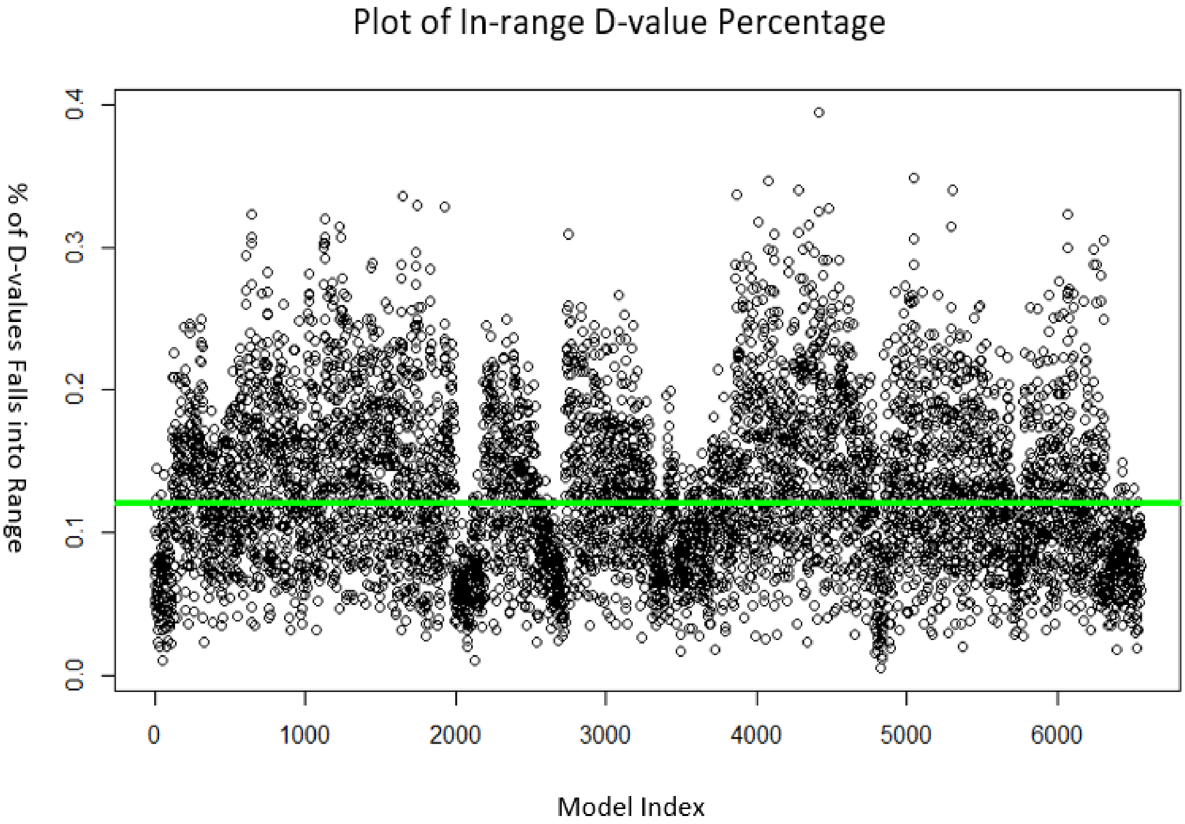
From the perspective of the median and mean, for all these models about 12% of their D-value falls into the range. For research purpose here, it is an acceptable level.

<u>R code to calculate in-range D-value percentage, summary table and plot</u>

```
rg<- function(x){
  ll=0.1
  ul=0.5
  sum(x > ll & x < ul)/length(x)
}

rg<- apply(dmat,1,rg)
plot(rg)
abline(h=0.12, col="green", lwd=4)
summary(rg)
```

Summary Table

Min	1st Qu.	Median	Mean	3rd Qu.	Max
0.005085	0.085593	0.121186	0.128409	0.164407	0.394915

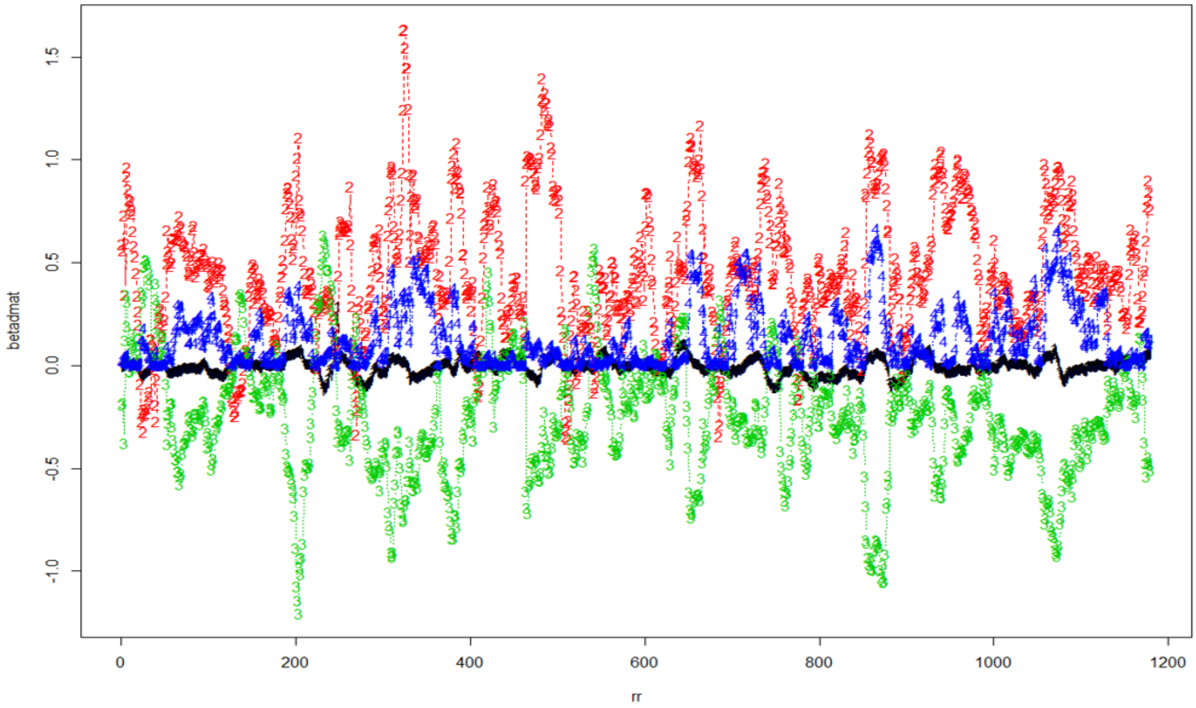


Graph 4.1.9 D-value in-range percentage of all models on the first scan

So, we will include 0.1 as the lower boundary for D-value in the following consistency calculation as one of the thresholds.

4.2 THRESHOLDS FOR CONSISTENCY

We are not only interested in the extremely good cases but are interested in when and how often the D values are high.



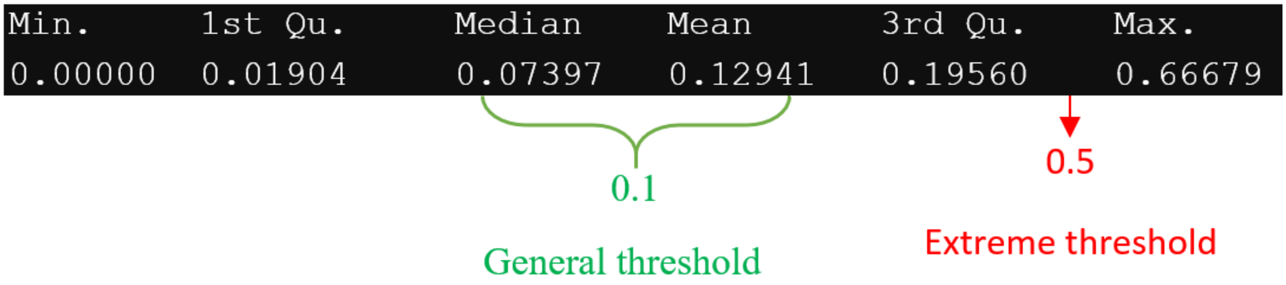
Graph 4.2.1 Plot of D-value, coefficients and constant item of model 4413 on the first scan

After observation on coefficients of variables in the above plot, it is not hard to see, when the coefficient of the first independent variable (Beta1, red points) is positive and the coefficient of the second independent variable (Beta2, green points) is negative, the D value at the same time point always tends to be high. In addition, when the difference between the two coefficients becomes larger, the D-value at that point tends to be even higher.

Now we know how to distinguish when the D-value is high, to solve the second question: how often the D-value appears to be large, another concept named "consistency" is introduced.

Consistency is calculated by summing up the number of time points where the D value is high, and coefficients are in accordance with the above conditions, then divided the sum by the total number of time points on one scan, thus we get a number less than 1. Consistency can be interpreted as a percentage which perfectly answers the question of how often the large D value appears.

As what was discussed in the previous part, setting the target threshold to 0.5 will only distinguish time points with an extreme situation happens occasionally. To represent a general situation, it is necessary to adjust the target threshold to a lower level. Threshold equals 0.1 is the middle point of mean value and medium, also it is the minimum requirement which is considered as qualified D value. With D-value less than 0.1, there is not a much significant improvement on a three-way



```
Min.      1st Qu.     Median    Mean      3rd Qu.    Max.
0.00000   0.01904     0.07397   0.12941   0.19560    0.66679
```

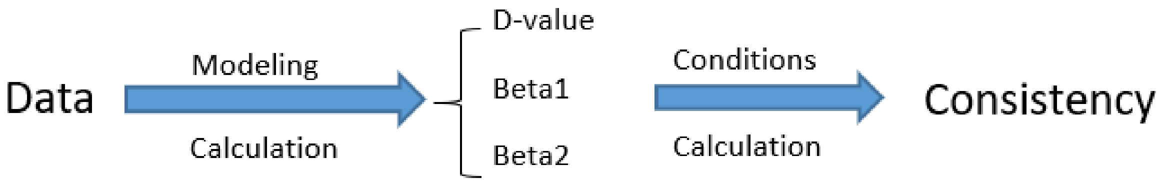
Graph 4.2.2 Model 4413 summary and threshold selection

The final condition for consistency is:

$$\left.\begin{array}{c} D > 0.1 \\ Beta1 > 0.1 \\ Beta2 < -0.1 \end{array}\right\} \Rightarrow \text{Consistent}$$

4.3 ALL MODELS TEST ON THE FIRST SCAN

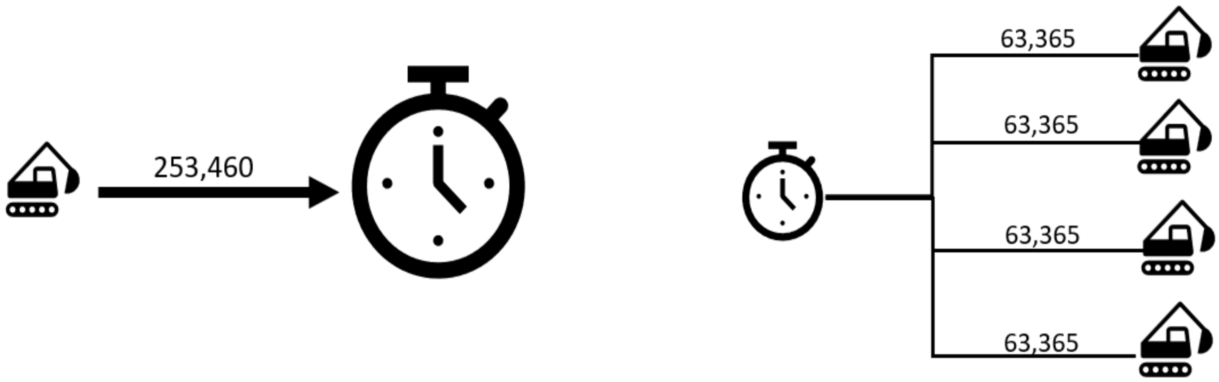
In the above process, for exploring and attempting purpose, the dependent variable is always set and fixed as a specific brain region. Based on all the works has been done so far, we established a system of calculation and judgment conditions. Now, it is feasible to start open restriction for the dependent variable and apply the system to all possible models and variables combinations.



In practice, the calculation involved in this process for 1,500 models on one scan typically needs two hours to be finished. The number of all possible variable combination is 253,460, and the total number of scans is 3,280. To test all of them one by one, it will take a large amount of time. Instead of testing all models on all scans, which is more comprehensive, test all models on the first scan is applied in this research. The strategy here is using the test on the first scan as a filter to select qualified models, and then redo the process and calculate consistency only for qualified models on all scans.

4.3.1 Parallel computing

Since the calculation is a time-consuming process, to improve time efficiency, parallel computing algorithm is implemented in R. This implementation calls all cores in CPU and instruct them to do the calculation simultaneously. For example, there are four cores available in one computer, the total time consumption will be decreased to ¼ of the original amount. As a result of parallel computing, it takes 85 hours in total to finish the consistency calculation for all of the 253,460 models on the first scan.

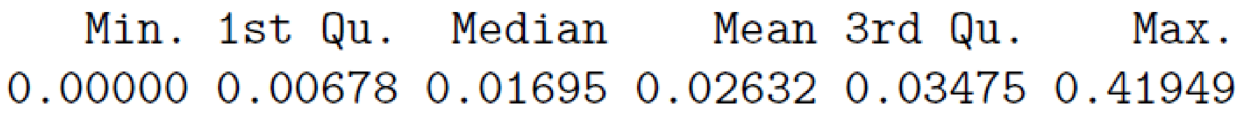


Note: the above calculation and time consuming is based on running R code on an ASUS N56 laptop with windows 10 64x system installed, the laptop has 4 cores and 8GB RAM. The programming structure used here is nested for loop.

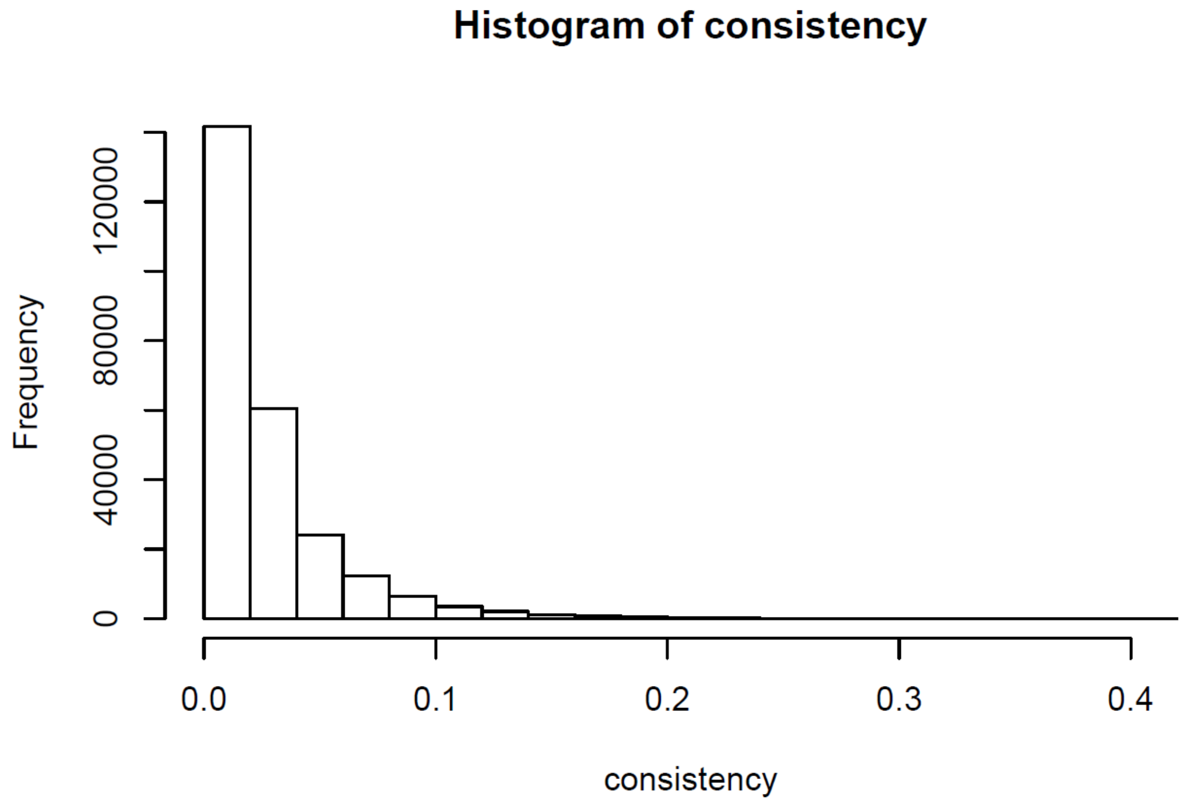
Time efficiency may be improved by applying more efficiently language or code, and running on higher performance computer. Cloud computation where more cores are available for calculation may significantly reduce the time consuming as well.

4.3.2 Qualified consistency threshold

The following are distribution summary and histogram which show details of the distribution of all models' consistency on the first scan:



```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.00000 0.00678 0.01695 0.02632 0.03475 0.41949
```



Similar to the previous threshold selection principle for D-value, we first investigate extreme good models. Here, I choose 0.35 as the threshold.

Of course, a different threshold can be used and will certainly return different results. For example, if the threshold is set to 0.30, around two hundred models will fall into "the bucket". With 0.35 as the threshold, only the top six models are selected, and it is relatively a tiny size group of models. Since I am only interested in top performance models here, 0.35 is good enough to be set as the qualified consistency threshold.

4.3.3 Qualified models

The top six models fall into "the bucket". They are:

Model No.	Dependent Variable (Z)	Independent Variable (X)	Independent Variable (Y)
4412	Precentral_L 2001	Occipital_Mid_L 5201	Occipital_Inf_L 5301
4413	Precentral_L 2001	Occipital_Mid_L 5201	Occipital_Inf_R 5302
10853	Precentral_R 2002	Occipital_Mid_L 5201	Occipital_Inf_L 5301
10854	Precentral_R 2002	Occipital_Mid_L 5201	Occipital_Inf_R 5302
64519	Frontal_Inf_Oper_L 2301	Parietal_Inf_R 6202	Angular_R 6222
230209	SupraMarginal_R 6212	Precuneus_R 6302	Cerebelum_Crus2_L 9011

4.3.4 Correlation review on pairwise brain regions

If we investigate the qualified models' variable combination in details, there are several brain regions appears repeatedly, dependent variable: region 1: Precentral_L, region 2: Precentral_R, and independent variable region 51: Occipital_Mid_L, region 53: Occipital_Inf_L, and region 54: Occipital_Inf_R.

As we all know, human brain regions have internal two-way relationships, correlation. The reason for these brain regions appears repeatedly could be an indicator of strong pairwise relationships.

The correlation between region1 and region2 is 0.7985, which is a very strong positive relationship. This explains why with the same independent variables, both combinations, with region 1 and region 2 as the dependent variable, has significantly similar performance because their signals are mutually boosted.

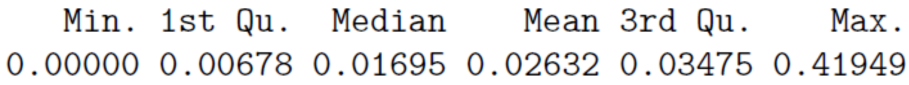
The same thing happens to region 53 and region 54, these two regions are assigned as independent variables in the different models, with other condition unchanged, no matter which of the two regions appear, the model is qualified. Like model 4412 and model 4413, model 10853 and model 10854, the only difference is switching an independent variable. The correlation between region 53 and region 54 is 0.7827, which is also a very strong positive relationship.

These pairwise relationships are existing when tested brain regions are located on the corresponding left and right side of the human brain. Although, we know that the left brain and right brain have their own different responsibility, in fact, the left brain is mainly responsible for logical thinking and the right brain is mainly responsible for feeling visualization, the above correlation phenomenon could still be a clue which indicates there exist more complex relationships and corporations between left and right.

4.4 SELECTED MODELS TEST ON ALL SCANS

4.4.1Acceptable consistency threshold

The acceptable consistency threshold is set to 0.04. Take the following summary table as an example:



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.00000 0.00678 0.01695 0.02632 0.03475 0.41949 
```

The 3rd quartile is 0.03475, which is close to 0.04. This is the true consistency distribution on many scans. Human brain activity is influenced by tons of external factors, even among four scans of one same person, it shows high volatility and difference in the original frequency data. A model who achieves consistency above 0.04 on thousands of scans stably cannot be coincident.

4.4.2 Consistency percentage

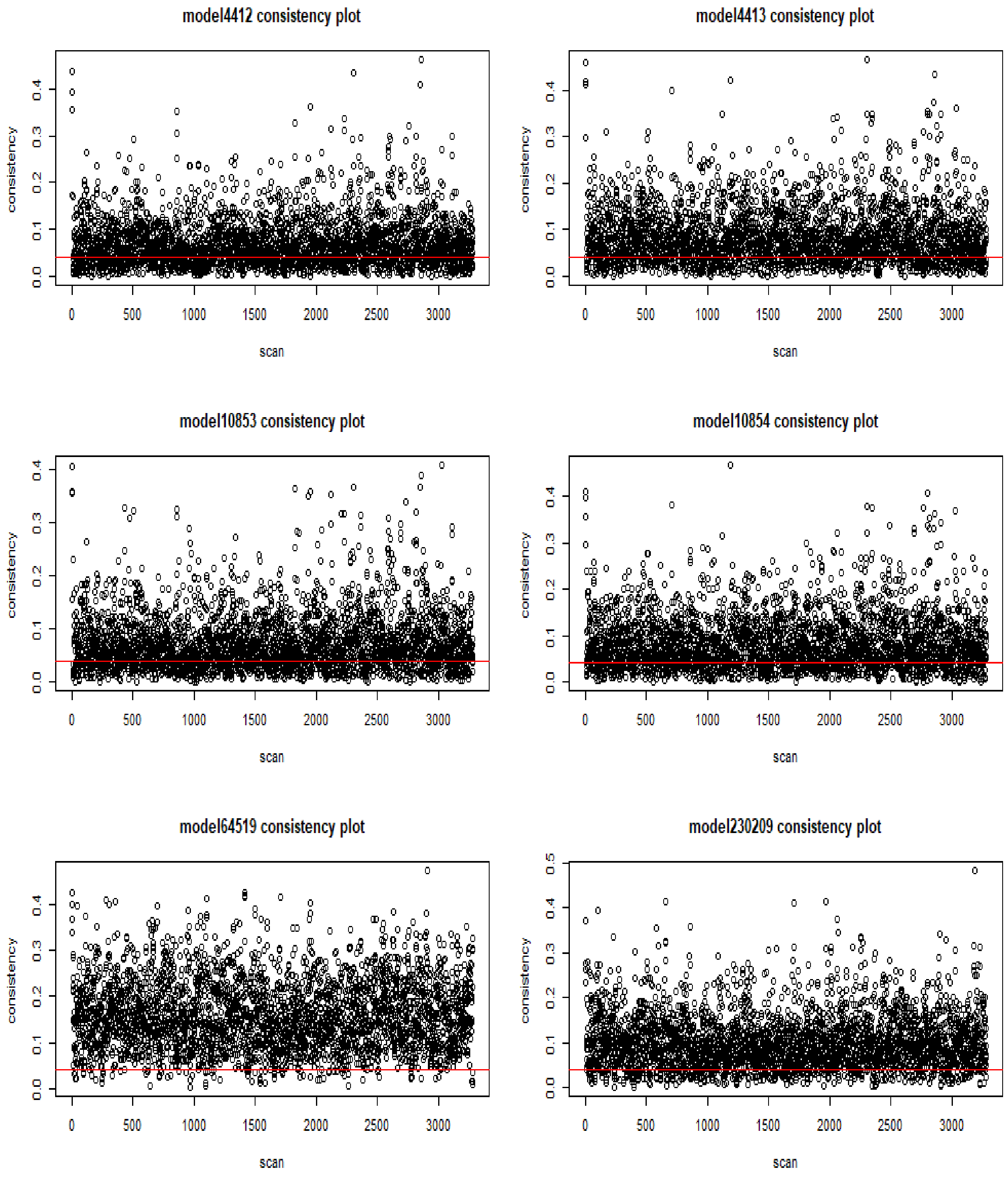
On all 3,280 scans, apply the same procedure to calculate D values and consistency for the six qualified models. Then compare the consistency of each model with an acceptable consistency threshold, add up the total number of scans where the consistency is greater than the acceptable consistency threshold, divide the sum by the total number of scans 3,280 to get a percentage. This percentage represents, on all scans, a model's overall performance. The percentage is named "consistency percentage". A high consistency percentage shows a model is stable on most of the scans.

4.4.3 The consistency percentage table

Model No.	Consistency percentage
4412	68.72%
4413	78.23%
10853	71.07%
10854	76.13%
64519	97.23%
230209	87.47%

Table 4.4.3 Consistency percentage of the six qualified models

All the six models return consistency percentage around or above 70%, this means these models have stable performance on at least 70% of the 3,280 scans.



Graph 4.4.1 – 4.4.6 Consistency plots for the six qualified models

4.5 Global signal and qualified models

> *The global signal is widely used as a regressor or normalization factor for removing the effects of global variations in the analysis of functional magnetic resonance imaging (fMRI) studies. However, there is considerable controversy over its use because of the potential bias that can be introduced when it is applied to the analysis of both task-related and resting-state fMRI studies.*

In order to further explore qualified model's sensitivity on the global signal, we control the global signal from the original data for regression and repeat the same calculations and analysis.

Because in the original data array, there is no global signal detected specifically, so we applied a simulation method to generate global signal data. The original data has 116 brain regions and 1,200 data points on each scan, for each data point, take the average for all the 116 brain regions, then subtract the average value from each of the original brain regions data to form a new data frame.

The new data frame is then inputted into the testing process, where the six qualified models are tested on all the 3,280 scans. Thus, we get the global-signal-controlled consistency matrices.

Apply the same general level standard of consistency on the global-signal-controlled matrices to evaluate the six models' performance, we get a new consistency percentage table below. The second column labeled "GB" is based on new global-signal-controlled matrices, and the third column labeled as "OG" is the original consistency matrices consistency percentage.

Model No.	Consistency Percentage (GB)	Consistency Percentage (OG)
4412	66.46%	68.72%
4413	72.32%	78.23%
10853	67.53%	71.07%
10854	68.26%	76.13%
64519	96.10%	97.23%
230209	68.60%	87.47%

Table 4.5.1

Compare global-signal-controlled consistency percentage with the original consistency percentage of the six qualified models

Comparing to the original consistency percentage, it is clear to observe that all the six models' consistency percentage decreased when we controlled the global signal. But the decreased portions are different for each qualified model, which indicates the models have different levels of sensitivity toward the global-signal.

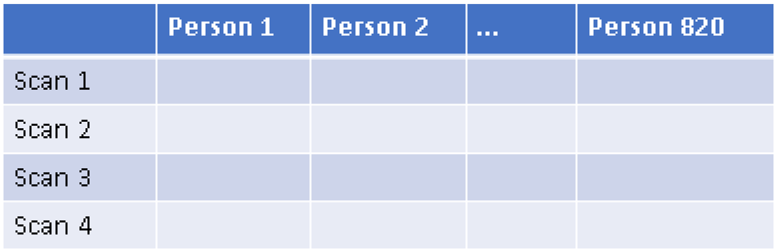
Model No.	Decreased Portion
4412	3.28%
4413	7.56%
10853	4.98%
10854	10.33%
64519	1.16%
230209	21.58%

Model 230209 has the most sensitive reaction toward global signal control, and model 64519, who has the highest consistency percentage among all other qualified models, has the least sensitivity level.

5 VARIANCE ANALYSIS ON SELECTED MODELS

Variance measures how far a set of numbers are spread out from their average value [2]. In this section, we will focus on discussion about consistency similarity for one model on the four scans of one person. Lower variance show that a group consistency is more concentrated, or in other words, closer to each other in terms of value. Higher variance represents the opposite situation which means the group consistency is more spread away from each other.

From the previous calculation and work, the consistency data is stored in matrices, one matrix for each model. For illustration purpose, below is a data table showing how a consistency matrix look like:



	Person 1	Person 2	...	Person 820
Scan 1				
Scan 2				
Scan 3				
Scan 4				

Graph 5.1 Appearance of a consistency matrix for one model

Each blank in the above data table contains a consistency value for each scan and each person. If we do column-wise variance calculation, we will be able to get consistency variance for each person.

5.1 OVERALL VARIANCE

Now we have a variance for each person, how to describe and decide whether the variance is small or large? It is a good idea to compare the column-wise variance with overall variance.

Overall variance represents how far the whole matrix of consistency are spread from their mean. If a column-wise variance is lower than the overall variance, it means compared to the whole set, this one column of numbers is more concentrated than the average level.

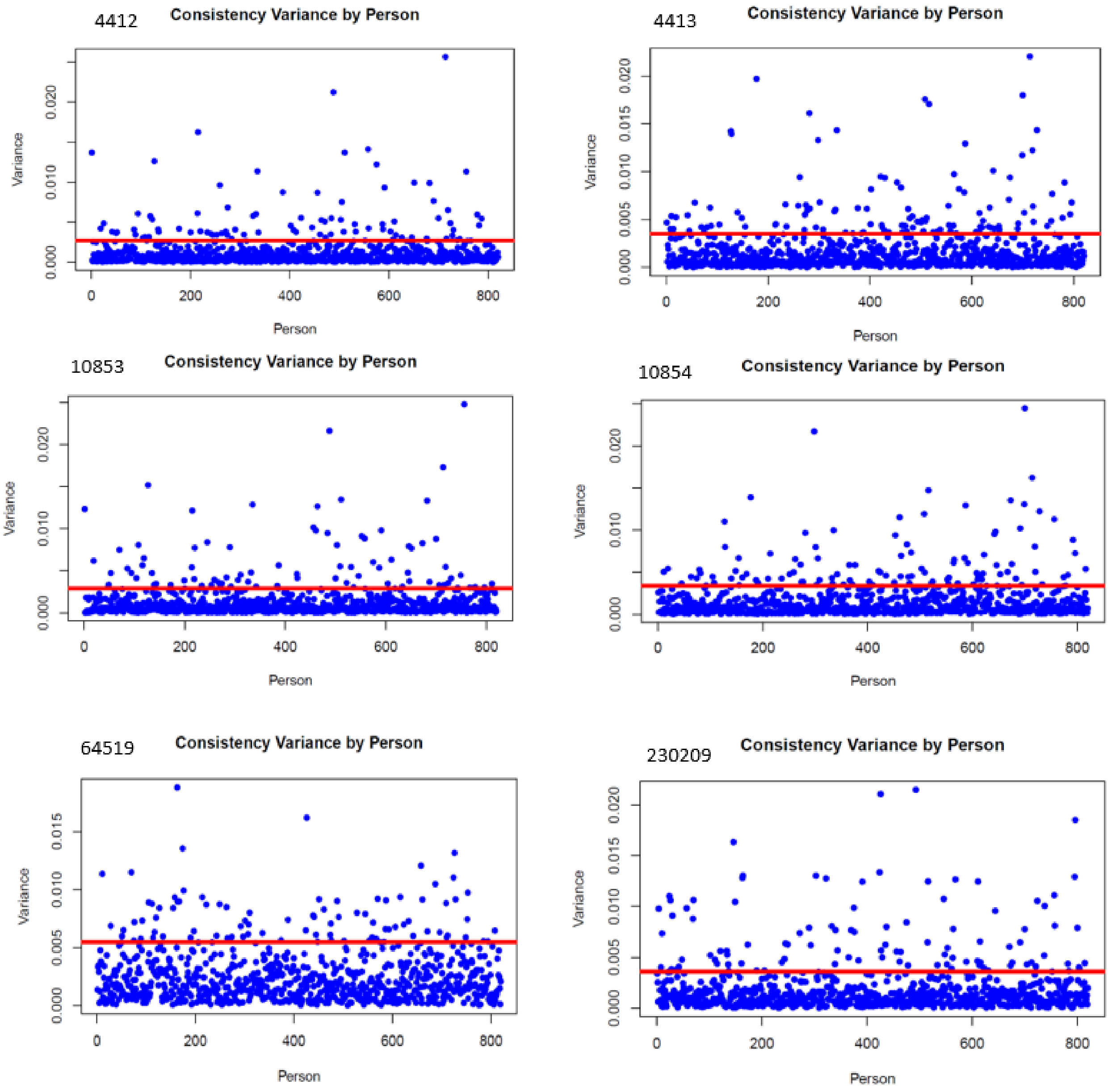
Overall variance is calculated by taking all consistency from the above matrix into consideration and using a simple function in R. The overall variance of consistency for each of the six qualified models we selected is shown in the following table:

Model #	Overall variance
4412	0.002697087
4413	0.003533631
10853	0.002910637
10854	0.003394082
64519	0.005453598
230209	0.003589780

5.2 CONSISTENCY SIMILARITY OF THE SAME PERSON

To determine how many people, have a similar consistency of their four scans, the variance of each person is compared to the overall variance of a model. If the personal variance is lower than the overall variance, the consistency of the four scans are similar, otherwise, they are not similar. Also, in order to clearly illustrate the level of similarity, a percentage is calculated by using the total number of people whose scans consistency are similar to divide the total number of people, 820, this percentage shows the portion of people who have similar consistency on one specific model.

Model #	**Percentage**
4412	89.15%
4413	86.10%
10853	89.63%
10854	86.95%
64519	89.02%
230209	87.56%



Graph 5.2.1-5.2.6 Personal consistency variance comparison with overall variance

Note: X-axis is person index, the y-axis is variance, red horizontal straight line is the overall variance, blue points are personal variance.

6 RANDOM EFFECTS MODEL

Human brain activity may be highly different from person to person. In the previous part, we looked into the consistency variance by person and found out that, for all of the six selected models, over 85% of people have similar or closed consistency in the four scans. To further analyze this phenomenon, a random effects model with a person as the only factor is established.

Random effects model is also called a variance components model. [3] Just like its name, this step is aimed at finding out how does "person" influence the variance of consistency.

Within-person variability is a measure of how much an individual tends to change in the sample. Specifically, within-person variability here measures the mean of the consistency change for the average individual change in the sample.

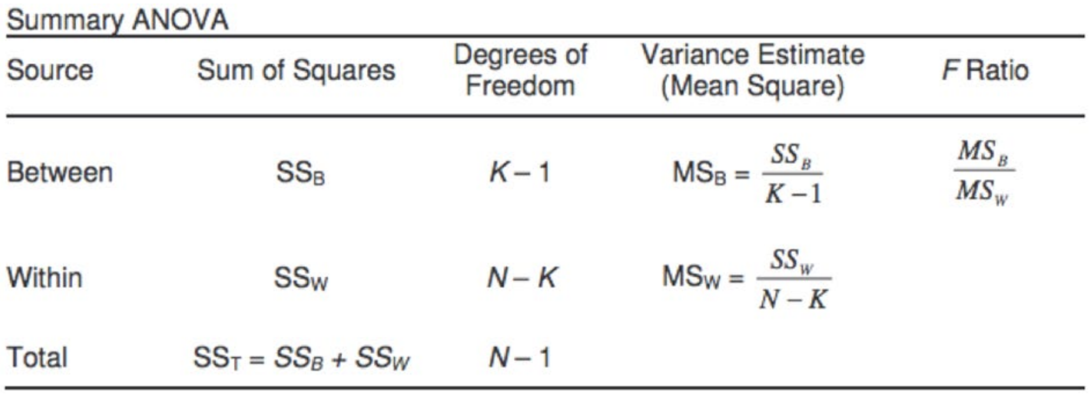
Between person variability measures the difference of the mean of consistency between individuals.

Total variability equals the sum of within-person variability and between-person variability.

The random effects model is built by taking consistency data from each of the six qualified models as the dependent variable, and person code 1 to 820 as a factor as the only independent variable. Then use linear regression models to build the random effects model.

To calculate within and between-person variability, ANOVA analysis is performed on this random effect model, the ANOVA table contains all results and numbers corresponding to the within and between-person variability.

For demonstration purpose, an example ANOVA summary table is shown as the following:



Summary ANOVA

Source	Sum of Squares	Degrees of Freedom	Variance Estimate (Mean Square)	F Ratio
Between	$\mathrm{SS_B}$	$K-1$	$\mathrm{MS_B} = \dfrac{SS_B}{K-1}$	$\dfrac{MS_B}{MS_W}$
Within	$\mathrm{SS_W}$	$N-K$	$\mathrm{MS_W} = \dfrac{SS_W}{N-K}$	
Total	$\mathrm{SS_T} = SS_B + SS_W$	$N-1$		

Graph 6.1 Demonstration of within and between-person variability in ANOVA table

Model No.	Total variance	Between	Within	Within/total ratio
4412	0.0221	0.0194	0.0027	12.17%
4413	0.0036	9.8745×10^{-5}	0.0035	97.28%
10853	0.0222	0.0194	0.0029	13.05%
10854	0.0057	0.0023	0.0034	59.36%
64519	0.0056	0.0001	0.0055	97.38%
230209	0.0044	0.0009	0.0036	80.80%

From the above data sheet, we can see that all six models have different level of variances, this makes it hard to compare one model with others. To better demonstrate the above results, we calculated within-person variability vs. total variability ratio. This clearly indicates for model #4413, #64519, #230209 and #10854 that they all have relatively high within/total ratio, which means in these models, person tends to change between scans over time. But for the other two models, #4412 and #10853, the variance does not heavily depend on the person, or in other words, individual tends not to change much over time.

7 RESULTS

- After testing all models on the first scan, there are six qualified models, the following list shows the components of each qualified model:

Model No.	Dependent Variable (Z)	Independent Variable (X)	Independent Variable (Y)
4412	Precentral_L 2001	Occipital_Mid_L 5201	Occipital_Inf_L 5301
4413	Precentral_L 2001	Occipital_Mid_L 5201	Occipital_Inf_R 5302
10853	Precentral_R 2002	Occipital_Mid_L 5201	Occipital_Inf_L 5301
10854	Precentral_R 2002	Occipital_Mid_L 5201	Occipital_Inf_R 5302
64519	Frontal_Inf_Oper_L 2301	Parietal_Inf_R 6202	Angular_R 6222
230209	SupraMarginal_R 6212	Precuneus_R 6302	Cerebelum_Crus2_L 9011

- All of the above six models reach around or above 70% consistency percentage.
- All of the above six models show consistency similarity within the four scans from the same person, and over 85% of the participated people display this kind of similarity.
- The random effect model with person as the only factor shows that: in model 4412 and model 10853, one person's brain activity of certain regions does not tend to have much change over time; in models 4413, model 64513, model 230209, and model 10854, one person's brain activity of specific regions tends to change much over time.

8 CONCLUSIONS

- The three-way model will always return more precise fit than the corresponding two-way models, however, the relative advantage of three-way models may not be stable during one scan, for one person. When chose qualified models to test on more scans, it is necessary to consider the overall performance, consistency instead of several extremely good cases.
- When evaluating a three-way model's performance, set an "acceptable threshold" for consistency and calculate consistency percentage. This indicates the overall performance of a tested three-way-model on all scans.
- Human brain activity highly depends on the individual.
- Some brain regions activity tends to change over time, other regions activity tends to be relatively stable or possible have certain frequency pattern.

9 LIMITATION AND FUTURE WORKS

The approach taken in this research is very computationally intensive. Hence, only some choices for various parameters were explored. The specific choices of those parameters were based on preliminary exploratory data analysis of time series data.

One of the future work directions is to calculate all models' consistency on each scan and summarize all the consistency values as well as classify models into different tiers by their overall level of consistency and other relevant model performance measurement.

The aim of this thesis is to propose a new perspective on the statistical research on human brain activity. The author believes there is tremendous space for new study and creativity in this field.

10 ACKNOWLEDGMENTS

This research was guided by Professor Peter Bajorski, I am thankful for his expertise that greatly supports this research, without his guidance, I would not finish the research and the thesis.

I am also grateful to Dr. Andrew Michael and Dr. Nathan Cahill for assistance with MRI imaging and data processing and precious feedback and comments.

To my fiancé Chaojie Yang, thank you for your encouragement, especially, you have been there for me through every crisis, and you are my endless source of great joy and love.

Last but not the least, I would like to thank my family: my parents and to my sister for supporting me spiritually throughout writing this thesis and my life in general.

Original data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

APPENDIX

Appendix A: Calculate D-value for one specific scan

```
# Calculate D-value for one specific scan


# Setup independent variable index combination
# with fixed response Z as region 1, we only need to generate for predictors
X and Y
cb_1 <- t(combn(c(2:116),2))

# Setup D-value matrix to store calculation results
h=20  # window width
m=1200-h # number of
k=(115*114)/2  # number of models/number of rows of D matrix
dmat1_1 <- matrix(data=NA, nrow = k, ncol=m)

# pull out data of the first scan first person from the array
p = 1
s = 1
scan <- Subset.Scans.arr[ , ,s,p]

# Calculate D-value and store into matrix using nested for loop


for (l in 1:k){
  for (i in 1:m){
    z <- scan[i:(i+19), 1]
    x <- scan[i:(i+19), cb_1[l,1]]
    y <- scan[i:(i+19), cb_1[l,2]]
    dmat1_1[l,i] <- summary(lm(z~x+y))$r.squared-
max((cor(z,x))^2,(cor(z,y))^2)
  }
}

# save the d value matrix as Rdata file
save(dmat1_4, file = "D:/data/dmatori4.Rdata")
```

Appendix B: Calculation and plotting of consistency and strength

```
####Consistency and strength###
consistency <- function(s,p,mi,rr,th,thd){
  scan <- Subset.Scans.arr[,,s,p]
  b0 <- vector()
  b1 <- vector()
  b2 <- vector()
  dv <- vector()

  betadmat <- matrix(data = NA, nrow = length(rr), ncol = 4)
  sa <- vector()

  for (i in (rr)){
    z <- scan[i:(i+19), 1]
    x <- scan[i:(i+19), cb_1[mi,1]]
    y <- scan[i:(i+19), cb_1[mi,2]]

    b0[i] <- summary(lm(z~x+y))$coefficients[1]
    b1[i] <- summary(lm(z~x+y))$coefficients[2]
    b2[i] <- summary(lm(z~x+y))$coefficients[3]
    dv[i] <- summary(lm(z~x+y))$r.squared-max((cor(z,x))^2,(cor(z,y))^2)
  }

  betadmat <- na.omit(cbind(b0,b1,b2,dv))
  sa <<- which(betadmat[,2]>0.1 & betadmat[,3]< (-0.1) & betadmat[,4]>thd,
arr.ind = TRUE)

  matplot(rr,betadmat,type = "b")
  print(paste("Consistency: ", (length(which(betadmat[,2]>0.1 & betadmat[,3]<
(-0.1) & betadmat[,4]>thd))/length(rr))))

  print(paste("Strength: " ,(length(which(abs(betadmat[,2])>th |
abs(betadmat[,3])>th))/length(rr))))
}



## Parameters:
# s <- 1 #Scan
# p <- 1 #Person
# mi <- 4413 model index, which model to plot
# rr <- c(1:1180) range
# th <- threshold to measure strength
# thd <- threshold to measure D values

# consistency(s=1,p=1,mi=4413,rr=c(1:1180),th=0.5,thd=0.1)
## if you want to check where in the scan is consistent
# abline(v=sa, col="purple")
```

Appendix C: Parallel Computing (R code)

```
# Function: Test all models on the first scan
test.fun <- function(mrange=1:100){
  load("cb.Rdata")
  load("scanmat.Rdata")
  con <- vector()
  for (m in mrange) {
    cnt = 0
    cb.mat<- cb[m,]
    for (i in 1:1180){
      mat <- scanmat[i:(i+19),]

      z <- mat[,cb.mat [1]]
      x <- mat[,cb.mat [2]]
      y <- mat[,cb.mat [3]]

      sum.obj<- summary(lm(z~x+y))
      v <- sum.obj$coefficients

      b1 <- v[2]
      b2 <- v[3]
      dv <- sum.obj$r.squared-max((cor(z,x))^2,(cor(z,y))^2)

      if (b1>0.1 & b2<(-0.1) & dv>0.1){
        cnt <-  cnt + 1
      }
    }
    con<-append(con, cnt/1180)
  }
  return(con)
}




#### Execution ####

setwd("D:/data/")
load("cb.Rdata")
load("scanmat.Rdata")

library("parallel")
source("potato3.R")
system.time({
  no_cores <- detectCores() - 1
  cl <- makeCluster(no_cores)
  con<- parSapply(cl,1:253460, test.fun )
  stopCluster(cl)
})
save(con,file = "models_consistency.Rdata")
```

Appendix D: Random Effects Model

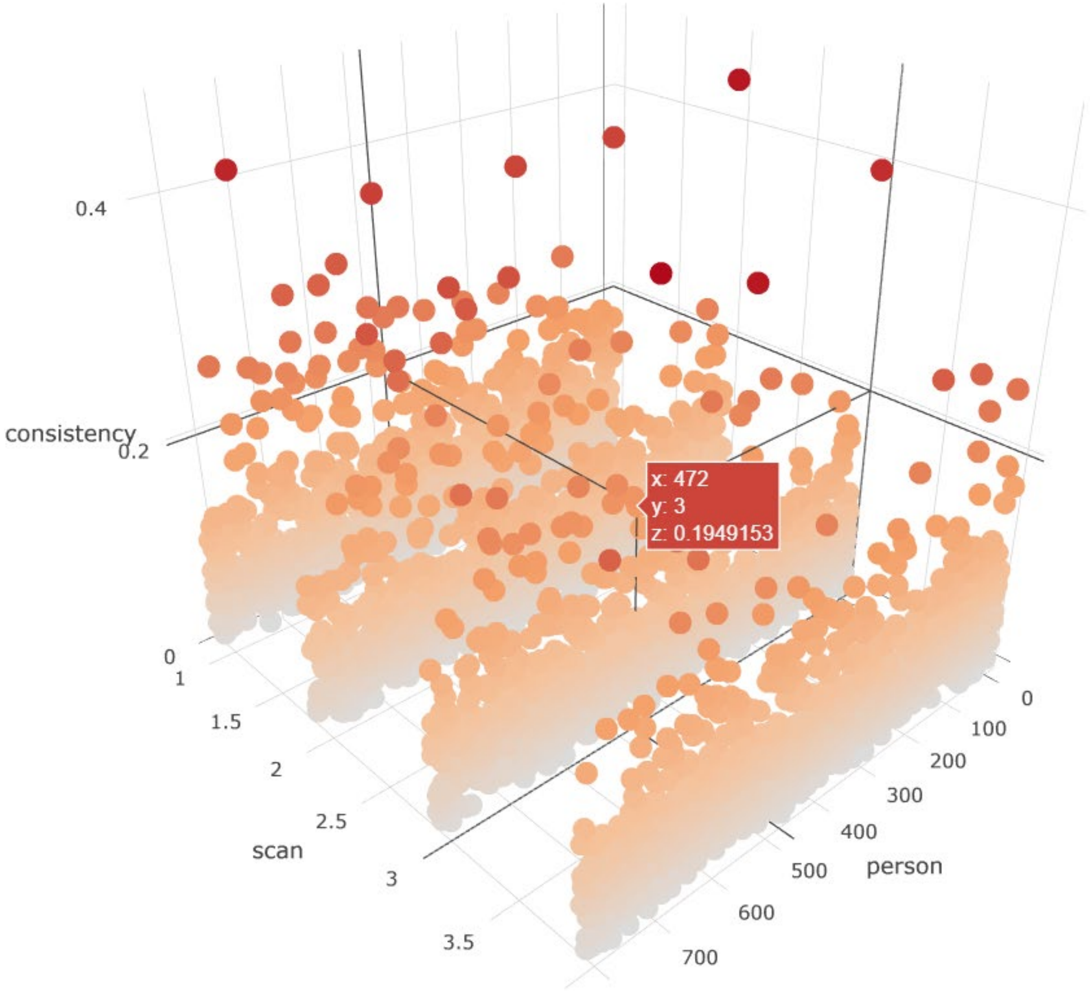
```r
# Random Effects model
factor <- as.factor(rep(1:820, each=4))
wtbt <- function(md=md4412){
  value <- as.vector(md[[1]])
  md <- as.data.frame(cbind(factor, value))
  obj <- lm(value~factor, data = md)
  av <- anova(obj)
  within <- av$`Mean Sq`[2]
  between <- av$`Mean Sq`[1]
  total <- within + between
  percentage <- within/total
  va <- cbind(within,between,total,percentage)
  print(paste("within person variance:", within))
  print(paste("between person variance:", between))
  print(paste("total variance:", total))
  print(paste("within/total percentage:", percentage))
  return(va)
}
```

Appendix E: Method to draw a 3-D interactive plot for model consistency (R code)

```r
# How to draw the 3-D interactive plot for model consistency?

# load and attach an add-on package "plotly"

library("plotly")


# load the consistency and strength data stored in lists

# for the specific model which you want to draw a 3-D interactive plot

load("md230209_2.Rdata")

# Transform and reorgnize the dataframe

df<- t(as.data.frame(md230209[[1]]))

df.v <- as.vector(df)

person <- rep(c(1:820),4)

scan <- rep(c(1:4),each=820)

df1 <- as.data.frame(cbind(person,scan,df.v))


# plot using plot_ly function

plot_ly(df1,x=person,y=scan,z=df.v,

  marker = list(color = ~df.v, colorscale = c('#FFE1A1', '#683531'),

showscale = TRUE)) %>%

  add_markers() %>%

  layout(scene = list(xaxis = list(title = 'person'),

                      yaxis = list(title = 'scan'),

                      zaxis = list(title = 'consistency'))

         )
```

Appendix F: The 3-D plot for model 4412 consistency

The following plot is shown as an example for interpretation purpose, because the documentation file does not allow insert of interactive dynamic plots, for all the six actual interactive 3-D plots, please check the attached HTML file named "*qualified models 3d plots*". Click on each static graph in the page, it will automatically direct to the corresponding interactive 3D plot.



REFERENCES

[1] Battistella, G., Najdenovska, E., Maeder, P. *et al. Brain Struct Funct (2017)* 222: 2203. *https://doi.org/10.1007/s00429-016-1336-4.*

[2] Steel, R. G. D.; Torrie, J. H. *(1960). Principles and Procedures of Statistics with Special Reference to the Biological Sciences.* McGraw Hill.

[3] Jump up^ Glantz, Stanton A.; Slinker, B. K. (1990). *Primer of Applied Regression and Analysis of Variance.* McGraw-Hill. ISBN 0-07-023407-8.

[4] Jump up^ Draper, N. R.; Smith, H. (1998). *Applied Regression Analysis.* Wiley-Interscience. ISBN 0-471-17082-8.

[5] "Correlation." *Categorical Data*. Accessed September 19, 2018.

http://www.stat.yale.edu/Courses/1997-98/101/correl.htm

[6] Liu TT, Nalci A, Falahpour M.*, The global signal in fMRI: Nuisance or Information?* Neuroimage. 2017 Apr 15;150:213-229. doi: 10.1016/j.neuroimage.2017.02.036. Epub 2017 Feb 16.

[7] "Variance." Wikipedia contributors. *Wikipedia, The Free Encyclopedia.* Wikipedia, The Free Encyclopedia, 26 Apr. 2018. Web. 27 Apr. 2018.

[8] "Random effects model." Wikipedia contributors. *Wikipedia, The Free Encyclopedia.* Wikipedia, The Free Encyclopedia, 26 Apr. 2018. Web. 15 May. 2018.