Estimacin de la edad de un cerebro mediante el anlisis de imgenes de resonancia magntica funcional

Moreno Durán, José Luis

20 July 2016

Contents

Chapter 1

Introduction

In this project we present a study of the brain connectivity and how this connectivity is affected by aging. This study has lead to defining several graph-based parameters that allow estimating the age of a given brain. Using these parameters, we propose a family of brain age estimators whose performance outperforms those of state-of-the-art techniques. We also present a tool set developed in C++ that has been designed specifically for this project and is able to perform complex operations such as brain segmentation, graph building from fMRI data and the computation of several graph descriptors such as clustering coefficient, average minimum path length, global efficiency or Small World coefficient.

1.1 Context

Functional magnetic resonance imaging (fMRI) was introduced to the medical field around 20 years ago. The fMRI is a functional neuroimaging procedure using MRI technology that measures brain activity by detecting associated changes in blood flow. This technique relies on the fact that cerebral blood flow and neuronal activation are coupled. When an area of the brain is in use, blood flow to that region also increases.

The primary form of fMRI uses the blood-oxygen-level dependent contrast. This is a type of specialized brain and body scan used to map neural activity in the brain or spinal cord of humans or other animals by imaging the change in blood flow (hemodynamic response) related to energy use by brain cells. Since the early 1990s, fMRI has come to dominate brain mapping research because it does not require people to undergo shots, surgery, or to ingest substances, or be exposed to radiation, etc. Other methods of obtaining contrast are arterial spin labeling and diffusion MRI.

Since it was introduced, the fMRI has been used in many brain related studies in both normal and pathologic brains. However, most fMRI studies aimed at characterizing brain activity in response to various active paradigms. In addition to this, strategies that aim to characterize the low-frequency oscillations of the ongoing fMRI signals when individuals are in a resting state are becoming more polular recently.

The datases obtained from a resting-state fMRI have been largery analyzed in the context of funcional connectivity and is also being used to evaluate more complex network features of the brain. These strategies have been applied to a number of different problems in neuroscience, which include important diseases such as Alzheimer's, schizophrenia, and epilepsy.

These deseases have a huge impact in today's society. Worldwide, nearly 44 million people have Alzheimers or a related dementia and only 1-in-4 people with Alzheimers disease have been diagnosed. The global cost of Alzheimers and dementia is estimated to be 605.000 million dollars [1]. Schizophrenia affects more than 21 million people worldwide, but only a half of them recieve treatment [2]. In addition, approximately 50 million people worldwide have epilepsy, making it one of the most common neurological diseases globally [3].

Resting-state fMRI based techniques have also been largely used in order to understand the natural aging of the brain. This is a very important goal to achieve. Understanding how evolves in time, in terms of its connectivity and modularity, can be very helpful in order to distinguish normal brains from brains that have a disease or can diseases in the future. Knowing how the human brain evolves with age, then, can be a very powerful tool to prematurely detect diseases and treat them even before they can actually affect the affect the patient's life.

The European Community has shown interest it is concentrating efforts in helping the research groups that work in this area. The Human Brain Project was recently created by the European Commission. This project aims to achieve a better knowledge of the brain by gathering efforts of different research groups and using high technologies as super-computing in order to help them [4]. In addition, organisations like the European Brain Council are also helping in the understanding of the human brain by helping to get more investment for the research groups and by educating the society on which are the main brain diseases and how can they affect to a person, and also on how to prevent them [5].

These are the main reasons that cause that the studies of brain structure and connectivity are increasing dramatically in the last years. Smith et al. 2011 [14] is one of our basic recent references when talking about Network modelling methods for fMRI. The brain can be modelled as a graph where vertices represent different neuronal regions and edges represent the way in these regions are related. Studies of the brain structure itself can help to detect present or latent brain illnesses. Other studies focus their efforts in understanding the way the age affects the neuronal structure which can be the key for understanding and preventing some age-related illnesses such as alzheimer. Our work is included in this second group and its main aim is to find a good way to predict the age of a brain network.

1.2 State of the Art

There are very different approaches on how the graph analysis can be used for studying the brain connectivity. Thompson et al. 2012 [8] point out several graph descriptors and measures that can have certain reliability when applied to neuronal networks represented by graphs. The reliability of these descriptors depends on the sparsity of the graph, which can vary depending on the threshold used when binarizing it. Here, we are discussing some of these descriptor; specifically, those that report results on age estimation.

Sala-Llonch et al. 2014 [7] divides the brain in 90 regions, compute the edges of the graph as the Person's correlation between the data of the regions and uses a threshold for binarising the graph. Then, they compute a set of descriptors, such as Clustering Coefficient or Average Minimum Path Length, in order to study the global and local connectivity of the neuronal network and how the descriptors and the edges evolve with the age. This work concludes that older participants showed lower connectivity of long-range connections together with higher functional segregation of these same connections. They also conclude that higher local clustering in older participants was negatively related to memory performance.

This work sets the basis of our study, as the reproduction of their results was our starting point. Results were obtained using the above descriptors upon a binarized graph formed by 15% of the positive, strongest weighted connections. From the reproduction of these results we corroborated that minimum value of the RMSE when predicting age was obtained with the Clustering Coefficient descriptor and its value was 8.5 years. Therefore, our main goal is to improve the quality of this estimation by decreasing the RMSE in our predictions.

Meier et al. 2012 [9] divide the brain in 100 regions and 4 functional networks, and use a Support Vector Machine classifier to classify the subjects into young or adult, focusing their efforts in studying the inter-network and intra-network connectivity without binarizing the graph. The results of this work claim that the positive weighted connections tend to strengthen with age while the negative weighted and long-range edges tend to get attenuated. They also conclude that the inter-network connectivity is strengthened with age whereas the intra-network connectivity tends to decrease.

Meunier et al. 2009 [10] also divide the brain in 90 regions but, using the wavelet correlation as weights of the edges, they binarize the graph and compute the centrality of each edge to determine the brain modularity in order to find if the subject is a young adult or and old adult. Other works, like Alexander-Bloch et al. 2010 [11], focus their efforts in studying parameters such as the modularity and how the different functional networks which are present in the brain structure are affected by diseases such as schizophrenia. Those works conclude that the global modularity of the brain remains very similar with age. However, the relation between some specific modules and functions of the brain do vary with age even when the global modularity remains the same.

1.3 Project goals

As stated before, in our study we take Sala-Llonch et al. 2014 [7] as starting point and try to improve their age estimation results. For achieving this goal, we study how the negative weighted connections impact in the behaviour of the descriptors, how well edges can separately predict the age and how the chosen descriptors behave when applied upon graphs formed by *reliable edges* which can predict the age of the brain with certain reliability. As the database that we used is relatively small, all our results were verified by performing a N-1 cross validation of the results. This validation produced very similar outputs to those obtained in the simple procedure. This way, we report a technique that is able to predict the age of the brain network with a minimum error of 7.6 years. These results imply and improvement of 10% in accuracy with respect to current state of art techniques.

A secondary goal of this project was to build a tool set that was able to read fMRI based data, build a graph from this data and perform several graph analysis operations upon the graph. The tool set that was developed during this project is able to achieve this goal, and allows the user to perform actions such as brain segmentation, graph building and prunning and computing several graph descriptors upon the resulting graph.

Chapter 2

Basic concepts used in this thesis

2.1 Undirectional weighted graphs

A graph G is a data structure which consists of two types of elements, namely vertices and edges. It is used to represent a set of elements (vertices) which are connected in some way (edges). A vertex is simply a node of the graph. The vertex set of G is usually denoted by V(G), or V when there is no danger of confusion. The order of a graph is the number of its vertices. An edge is a set of, or a connection between, two elements or vertices. The edge set of G is usually denoted by E(G), or E when there is no danger of confusion. The two endpoints of an edge are also said to be adjacent to each other.

A weighted graph associates a label (weight) with every edge in the graph. Weights represent the value or strength of the connection between two vertex of the graph and are usually represented by real numbers. They may be restricted to rational numbers or integers. It is said that a graph is undirectional when, for each connection between two given nodes, the weight from the first node to the second is equal to the weight from the second node to the first.

2.2 Pearson's correlation

Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. It can be computed according to the next equation.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{2.1}$$

2.3 Data structures and basic algorithms

2.3.1 Auto-balanced binary search tree AVL

A binary search tree, BST, is a tree shaped data structure used in computer science. This structure is defined by the property that, for each subtree and a given comparison function for comparing nodes, all the elements of the left child have a lower value than the root, while the elements of the right child have a greater value than the root.

In 1962, Georgii Adelson-Velskii and Yvegeniy Landis invented the first auto-balanced binary search tree. This new data structure, known as AVL tree in honor of its inventors, is an improvement of the classic BST. The main trait of the AVL tree is that it is an auto-balanced tree. This means that the height of the left branch will never be greater than the height of the right branch plus one, and viceversa. Due to that, the AVL tree has the great advantage that it allows searching, inserting or deleting elements of its structure with a $O(\log n)$ computational cost, with n being the input data volume. This means a great efficiency improvement in applications that require constant use of such kind of operations.

In a graph segmentation process, it is very common having to search, delete or insert new edges to the graph during the nodes fusion phase. For this reason, the AVL tree is the ideal structure for storing the data related to a graph in this kind of processes.

2.3.2 Distances matrix

A distances matrix is a matrix whose cells represent the distances between the elements represented by its indexes. In this context, we define the distance between two nodes as the shortest path between both nodes. The shortest path between two nodes is defined as the minimum sum of edge weights between both nodes. In an unweighted graph, the distances between two nodes is just the minimum number of edges between both nodes. The distances matrices are, then, NxN sized symmetrical matrices, with N being the number of elements whose distances are being computed.

For any kind of graph, the distances matrix is the fastest and easiest structure for obtaining the distance between any pair of vertices or even between a vertex and the rest of the graph. It is only needed to access one position for obtaining a single distance, and the row related to an element for obtaining its distance with the rest of the graph.

2.3.3 Dijkstra's algorithm

The Dijkstra's algorithm, also known as shortest paths algorithm, was first described in 1959 by Edsger Dijkstra. This algorithm computes the shortest distance between the selected vertex as the origin and the rest of vertices of the graph.

The algorithm explores every possible path and selects in each step the node with lower distance, that is, the edge of the current node wich has the lowest value. This operation is repeated until the algorithm finds the shortest paths between the origin node and the rest of the graph. If the graph is unweighted, the algorithm just finds the minimum number of edges between the origin node and the rest of the nodes. This algorithm, however, does not work with graphs which have negative weighted edges because it can exclude future iterations that could possibly lower the total path length when adding negative values.

The Dijkstra's algorithm has a $O(|V|^2)$ complexity, with V being the number of vertex of the graph. This cost can be improved slightly improved by using priority queues, obtaining a final cost of $O((|E| + |V|) \log |V|)$. The algorithm, written in pseudocode, can be found ahead.

```
function Dijkstra (Graph G, output_node s)
  //A vector is used for storing the distances from the output_node to the rest
  int distance[n]
  //Initialization of the node
  boolean visited[n]
  //boolean vector for controlling the vertices of which we already have the minimum distance
for each w from V[G] do
     If (no edge exists between s and w) then
         distance[w] = Infinite
     Else
         distance[w] = weight(s, w)
     end if
  end for
  distance[s] = 0
  visited[s] = true
  //n is the number of vertices of the graph
  while (not_all_nodes_visited) do
     vertex = get_mimimum_from_distance_vector that has not been visited;
     visited[vertex] = true;
     for each w from successors(G, vertex) do
         if distance[w]>distance[vertex]+weight (vertex, w) then
            distance[w] = distance[vertex]+weight (vertex, w)
         end if
     end for
  end while
end function.
```

Algorithm 1 *Dijkstra's algorithm without priority queues.*

2.3.4 Bubble sort algorithm

The bubble sort algorithm is one of the simplest algorithms for sorting one-dimensional vectors, which can contain from simple data types to complex objects. The algorithm consists in scrolling through the vector that is being sorted and comparing each element with the following. If the elements compared are not in the correct order, the algorithm switches its positions and continues scrolling. The algorithm needs to scroll through the vector several times until it is completely ordered.

In the worst case, the bubble sort algorithm has a $O(n^2)$ cost, with n being the size of the vector that is going to be sorted. Due to it's simplicity, this algorithm is a good option for sorting short or partially ordered vectors, which are the cases in which it is used in this project. The algorithm, written in pseudocode, can be found ahead.

```
function BubbleSort( A : array of n elements indexed from 1 to n)
  for i from 1 to n-1 do: //n-1 scrollings
       for j from 1 to n-i do: //the scrolling
         if A[j] > A[j+1] then //if the elements are not in order
           switch A[j] and A[j+1] //they are switched
       end for
  end for
end function
```

Algorithm 2 *Bubble sort algorithm.*

2.4 Estadistical analysis: linear regression and P-value

In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable and one or more explanatory variables. In linear regression, data are modeled using linear predictor functions, and unknown model parameters are estimated from the data.

The p-value is a function of the observed sample results that is used for testing a statistical hypothesis. Before the test is performed, a threshold value is chosen, called the significance level of the test, traditionally 5% or 1% and denoted as α.

If the p-value is equal to or smaller than the significance level, it suggests that the observed data are inconsistent with the assumption that the null hypothesis is true and thus that hypothesis must be rejected. However, this does not automatically mean the alternative hypothesis can be accepted as true.

2.5 Graph descriptors

The main graph descriptors that were used throught this thesis are presented in this section. These descriptors are: Clustering Coefficient, Average Minimum Path Length, Global Efficiency and Small-World coefficient.

2.5.1 Clustering coefficient

In graph theory, a clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. Evidence suggests that in most real-world networks, and in particular social networks, nodes tend to create tightly knit groups characterised by a relatively high density of connections; this likelihood tends to be greater than the average probability of a tie randomly established between two nodes.

Two versions of this measure exist: the global and the local. The global version was designed to give an overall indication of the clustering in the network, whereas the local gives an indication of the embeddedness of single nodes.

$$CC = \frac{1}{n} \sum \frac{2t_i}{k_i(k_i - 1)} \tag{2.2}$$

where t_i is the number of links between the neighbours of the i region and k_i is the number of neighbours of the region i.

2.5.2 Average minimum path length

Average minimum path length is a concept in network topology that is defined as the average number of steps along the shortest paths for all possible pairs of network nodes. It is a measure of the efficiency of information or mass transport on a network.

$$AMPL = \frac{1}{n} \sum \frac{\sum d_{ij}}{n - 1} \tag{2.3}$$

where n is the number of regions and d_{ij} is the distance between the regions i and j.

2.5.3 Global efficiency

The efficiency of a network is a measure of how efficiently it exchanges information. The concept of efficiency can be applied to both local and global scales in a network. On a global scale, efficiency quantifies the exchange of information across the whole network where information is concurrently exchanged. The local efficiency quantifies a network's resistance to failure on a small scale. That is the local efficiency of a node i characterizes how well information is exchanged by its neighbours when it is removed.

$$GE = \frac{1}{n} \sum \frac{\sum \frac{1}{d_{ij}}}{n - 1} \tag{2.4}$$

where n is the number of regions and d_{ij} is the distance between the regions i and j.

2.5.4 Small-World coefficient

A small-world network is a type of mathematical graph in which most nodes are not neighbors of one another, but most nodes can be reached from every other by a small number of hops or steps. Specifically, a small-world network is defined to be a network where the typical distance L between two randomly chosen nodes grows proportionally to the logarithm of the number of nodes N in the network. The Small-World coefficient is used to determine whether a network is a small-world network or not.

$$SW = \frac{\frac{C}{C_{rand}}}{\frac{L}{L_{rand}}} \tag{2.5}$$

where C is the number Clustering Coefficient of the graph, C_{rand} is the Clustering Coefficient of a random network with the same number of regions, L is the Average Minimum Path Length and L_{rand} is the Average Minimum Path Length of a random network with the same number of regions.

$$C_rand = \frac{k}{N} \tag{2.6}$$

$$L_rand = \frac{\log(N)}{\log(k)} \tag{2.7}$$

where N is the number of vertexes of the graph and k is the average number of edges of each node.

2.6 Graph segmentation

The graph segmentation is a technique that consists in creating a hierarchical graph from a weighted graph. The resulting graph shows how the regions from the original graph group into bigger similarity regions, given that in each iteration of the segmentation process the two regions with the most similarity are grouped into a single region. Taking in consideration the structure of the segmentation process, the Binary Search Tree (BST) data structure is perfect for storing the hierarchical graph. The leafs of the tree represent the original graph and each level of the tree represent a fusion between two nodes.

For building the hiearchical graph, it is very important to choose how the regions are compared when two regions are going to be merged. In this project, when merging two original regions the algorithm allows to choose between the weight, the distance, or a combination of both as the similarity function. However, when a new region is created as a result of one or more previous fusions, there relationship between the new region and the rest can also be chosen. More specifically, the algorithm allows to chose the relation between the new region and each of the rest. The weight of the new region with the rest can then be calculated as the mean of the weights or distances of the original regions that form it, or directly as the weight of the closest original region that forms the new region to the region we are computing the similarity function with.

Chapter 3

Studies performed

3.1 Graph analysis of brain structure

3.1.1 Creation of a graph from the fMRI data

In our study we used the same database used in Sala-Llonch et al. 2014 [7]. One hundred and four healthy older adults (mean age: 64.87, Standard deviation [SD]: 11.8; 56 females, 48 males) were included in the study. Six individuals were excluded a posteriori due to vascular sub-cortical lesions or abnormal cognitive performance, leaving a final sample of n=98.

All participants were scanned with a 3T MRI scanner. The scanning protocol included functional MRI acquisition during a 5-minute resting-state and a high-resolution 3D structural dataset. For the resting-state fMRI, participants were asked to close their eyes, not to fall asleep, and not to think about anything special. Functional datasets from resting fMRI were preprocessed individually. Preprocessing was carried out using tools available in FSL software. Briefly, it included the removal of the first five scans, motion correction, skull stripping, grand mean scaling, and temporal filtering [7].

The Automated Anatomical Labeling (AAL) atlas [7] was used to parcellate the whole brain into a set of Regions Of Interest (ROIs). The AAL atlas includes 45 ROIs in each hemisphere and is based on anatomical landmarks on the standard MNI surface. AAL regions were registered to each individual functional space using previously obtained transformation matrices in order to extract ROI-associated time series and to construct networks of functional connectivity.

We computed the connectivity of each pair of ROIs using the Pearson's correlation between the associated time series, obtaining a complete weighted graph that represents how the different ROIs are related. In this resulting graph, the ROIs of the brain represent the vertexs (90 in total), while the correlations between them represent the edges (4005 in total, since the initial graph is built completely).

3.1.2 Correlation of edges with age

The first study that we performed consisted in seeing how each edge of the graph, that is, how each relation between ROIs correlates with the age of the subjects. We wanted to see if any of the 4005 connections would be a good predictor of the age by itself. For doing so, a vector with the 98 values of each subject was obtained for each edge of the graph and we computed the correlation of each of these vectors with the age vector. For each of these correlations, we measured its P-Value, its Root Mean Square Error (RMSE) and its slope, R.

Before starting with the experiment, we detected a defective subject. By simply plotting the vectors of each edge, we discovered that one of the subjects, with age 81, presented atypical values: very similar values for all edges which were also very different from the values of the rest of the subjects, including those with similar age. Therefore, we decided to exclude this subject from further studies and to continue with the other 97 subjects.

After computing the correlation of each edge of the graph with the age, we made a first classification of the edges. Those edges whose the P-Value of the correlation with the age was greater than 0,05 were considered as 'reliable edges' and the rest were considered as 'non-reliable edges'. We obtained a total of 762 reliable edges, meaning that each of these edges by itself have a certain capacity of predicting the age. This group of reliable edges was also divided into two different subgroups depending of the sign of the slope of its correlation with the age. The regression's slope of 443 of the edges was negative, meaning that their value (the Pearson's correlation between its nodes, which we will call *rho* from now on) decreases as the age increases. The rest, 319 edges, had a positive slope, meaning that their *rho* value increases with the age.

After analysing the RMSE value of reliable edges we found that they are capable of predicting the age with error values from 8.47 years to 9.9 years. As a result of this study, we conclude that there are some edges that have certain capability of predicting the age of the brain separately. The best result found, shown at figure 3.1, is the edge that connects the Superior Parietal Right region with the Angular Right region, with a prediction error of 8.47 years, whose correlation has a positive slope .

3.1.3 Graph descriptors

After studying each edge separately, the next step of our study was to study the graph which represents the brain as a whole, and to see if some of the descriptors which represent a graph can be a good age predictor. We want to start with simple descriptors so we have to binarize the graph. For doing this binarization, we applied different thresholds to the graph for keeping only those edges whose weight we consider more relevant in each case.

It is well known that the human brain is characterized for having a Small-World topology [12]. A Small-World network is characterized for having high Clustering Coefficient (CC) values and low Average Minimum Path Length (AMPL) values. A Small-World network will always have a Small-World co-

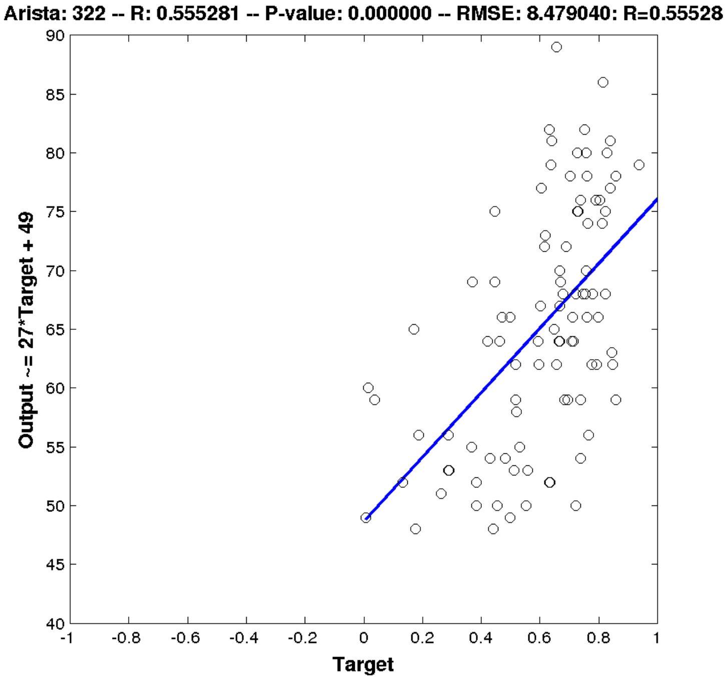


Figure 3.1: Regression of the connection of the Superior Parietal Right region with the Angular Right region .

efficient higher than 1.0. Having this into consideration, we decide to take these two descriptors as the starting point of our study.

However, when applying thresholds to the graph edges for binarizing we can obtain disconnected graphs as a result. This causes that the AMPL descriptor loses its sense since there will be pairs of vertices that will not have a path connecting them, causing infinite distances to appear[11] which result in values of AMPL equal to infinite too. For dealing with this problem we decide to include the Global Efficiency (GE) descriptor, which measures the efficiency of the graph [13] and is computed taking into account the inverse value of the distance between vertices. For connected graphs, the GE descriptor is the inverse of the AMPL descriptor. In disconnected graphs, those vertexs which are disconnected will have a contribution equal to zero in the descriptor, solving the problem that we had with AMPL.

Therefore, in the next sections our study will be based on seeing the behaviour of the descriptors already mentioned (CC, AMPL and GE) when computed upon binarized graphs obtained by applying threshold values to the main graph. We decided to keep AMPL despite its problems with disconnected graphs because, even with these problems, it is still a widely used descriptor. We will also study the evolution of the Small-World coefficient in the different scenarios. All the descriptors and coefficients are computed as stated in Rubinov et al. [6]

3.1.4 Impact of the negative edges in the graph structure

The processing of the data obtained from an fMRI (software and tools used, movement corrections, filtering, elimination of undesired areas, use of masks, etc.) and the methods used for defining or labelling the nodes play an important role when creating the data that will be stored in each vertex of the future graph representing an individual. Due to that, the mentioned pre-processing also affects to the correlations between the vertices of the graph, since they are directly calculated using the information contained in each of the vertices.

When using the Pearson's correlation to generate the edges of the graph, some correlations with negative value appear. There is an open debate on whether these negative correlations indicate the existence of some kind of antiphase relationship between the regions of the brain cortex that they represent or if, conversely, these negative values are nothing else that noise or residual values generated by the data pre-processing techniques. Some works even suggest that edges with negative weight should not be taken into consideration when computing graph descriptors based on brain structures [6] [7].

Behaviour of the descriptors depending on the weight sign

For bringing our own conclusions to this debate we tested how the three descriptors that we are using behave depending on the chosen correlation threshold used to binarizing the graph. Each descriptor has been tested in three different scenarios. In each scenario, the P-value of the linear regression between the 97 samples of each descriptor and the age vector is computed, and the graph used

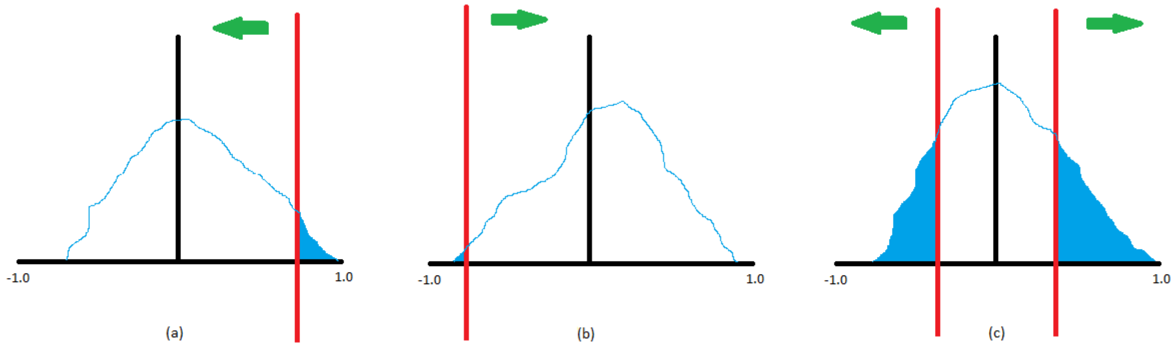


Figure 3.2: Threshold used in the scenarios 1, 2 and 3, respectively.

in each scenario is computed with a different kind of threshold. Therefore, a descriptor is considered to be statistically relevant for a given threshold value when its P-Value is lower than 0.05.

- Scenario 1: The specified percentage of edges whose value is greater than the threshold are kept. This scenario is used to study mainly the positive correlations. Figure 3.2-a.
- Scenario 2: The specified percentage of edges whose value is lower than the threshold are kept. This scenario is used to study mainly the negative correlations. . Figure 3.2-b.
- Scenario 3: This scenario uses a symmetric threshold which binarize the graph using directly the absolute value of the edges. Figure 3.2-c.

The third scenario, which covers positive and negative weights, shows a similar behaviour than the first scenario, which only covers positive weights. It can be concluded then, that negative weights have little impact and that they are related to noise or residual data caused by the processing of the fMRI data, and not represent a real connection between regions of the brain cortex. Analysing the second scenario, it can be observed that the AMPL misbehaves in general. Some P-values below 0.05 appear around a threshold value of 20%, but the curve never drops below 0.001 unlike what we observe in the first scenario. Comparing the minimum values obtained in both scenarios 1 and 2, it is noticed that the minimum values of the first scenario are four orders of magnitude lower than those obtained in the second scenario. This means that the positive correlations have a stronger meaning than the negative ones for the AMPL descriptor. Talking about the CC descriptor, it shows P-values below 0.001 for thresholds between 48% and 78%. However, for this percentages range, the threshold starts to keep not only negative correlations, but also positive correlations whose value is near to zero. It is considered, then, that this descriptor also misbehaves in the second scenario.

Sign of the weight of the reliable edges

In addition, we decided to study the sign of the weight of the edges marked as reliable. This study could be reliable because if the edges that are good predictors of the age had a negative weight, that would mean that negative weighted edges could bring some valuable information to the graph.

Therefore, we decided to plot the histograms of the weights of the reliable edges for each of the 97 subjects. The results of this experiment show that the reliable edges whose correlation with age has a positive slope also have a positive weight in most of the cases. On the other hand, the reliable edges whose correlation with age has a negative slope show histograms whose shape is similar to a normal distribution centred in 0, which mean that this kind of edges can have either negative, neutral or positive weights, and it is the evolution of its weight, and not the sign, which makes this edges good predictors of the age.

Taking this into consideration, we can conclude that edges with positive weight do have valuable information. On the contrary, the importance of the edges with negative weight is compromised once more because we can't stablish a clear correlation between the negative sign of an edge's weight and the sign of the slope of its correlation with the age.

After both experiments it can be concluded that negative weights have little importance and that they are rather noise or residual data caused by the processing of the fMRI data, and not represent a real connection between regions of the brain cortex. In the next chapter of the study we will focus only on the positive weights, and the thresholds applied will be the same as we exposed in the scenario 1 of this chapter.

3.1.5 Correlation of graph descriptors with age

After the previous experiments, we decided that the next step was to study the behaviour of the three descriptors (CC, AMPL, GE) computed upon a graph formed exclusively by the reliable edges found at the section 2.1. More specifically, we computed the descriptors upon three different graphs: a graph with all the reliable edges, a graph with the reliable edges whose correlation with the age has a positive slope and a graph with the reliable edges whose correlation with the age has a negative slope. We decided to compute the descriptors in these three scenarios because all of them are graphs formed by reliable edges, and we found reliable to find if a structure built with such kind of edges can be a good predictor as a whole.

We also computed the Small-World coefficient to analyse whether the subgraphs that we used at every experiment kept their Small-World network structure or not. At every experiment, we binarized the graph using a threshold which removed the edges with weight lower than its value and set the rests weight to one. We used a scrolling for the threshold, with its values varying from 0.0 to 1.0 in steps of 0.01, always removing the edges with negative weight. Therefore, for each value of the threshold we obtained a vector of 97 values for each descriptor and we computed the correlation of these three vectors with the age, obtaining its P-Value and Root Minimum Square Error. In the figures of this section we can see the P-Value and the RMSE obtained for each threshold for each of the three graphs.

Reliable edges graph

In this first case, figure 3.3, we can observe that the coefficient which better predicts the age is the Clustering Coefficient. This descriptor presents a minimum RMSE value of 8.4 years, which is slightly better than the result obtained by the best edge of the section 2.2, while the AMPL and GE descriptors show a minimum RMSE of 8.7 years. It is important to point that for every descriptor the minimum values of RMSE belong to sections of the threshold that have also a P-Value lower than 0.05. Also, in this case we always maintain a value of the Small-World coefficient higher than one, which means that the graph we are working with maintains the Small-World network structure.

Positive reliable edges graph

In the case of the graph formed by the reliable edges whose correlation with the age has a positive slope, figure 3.4, we observe that the Global Efficiency is the best descriptor, with a minimum RMSE of 8.6 years. The CC and AMPL descriptors show a minimum RMSE value of 9.2 years. Like in the previous scenario, the minimum values of RMSE belong to sections of the threshold where the P-Value is lower than 0.05, and the Small-World coefficient is always higher than one.

Negative reliable edges graph

This last scenario, shown at figure 3.5, is specially interesting. In the figures, we can see that the CC and AMPL coefficients show minimum values of RMSE of 9 and 9.3 years, respectively, which are not specially interesting. However, the Global Efficiency descriptor shows a minimum RMSE value of 7.6 years, which means that the GE descriptor is the best age predictor that we found in all of our experiments computed in a graph formed by the reliable edges whose correlation with the age has a negative slope. This minimum value also belongs to a section of the threshold where the P-Value is lower than 0.05. However, in this case we find that the Small-World coefficient is lower than 1.0, which means that the graph used in this escenario is not a Small-World network. We discuss what does that mean in the next section.

3.1.6 Small World results analysis

As we mentioned before, it is well known that human brain networks follow a Small-World network structure [12] [13]. That means that the brain networks are characterized for having a high Clustering Coefficient and a low Average Minimum Path Length values. This is possible due to the fact that the brain has short distance connections that make possible a high CC value, and long distance connections that make possible a low AMPL value between all regions.

In addition, we also know that the long distance connections of the brain tend to fade with the age, which means that the weight of long distance connections tends to decrease with age [7] [9]. Therefore, we could expect that the edges whose correlation with age has a negative slope correspond with long distance connections of the brain.

In the previous section we have seen that the graph formed by all the reliable edges had a Small-World coefficient value higher than 1.0. We could expect then, that this graph is also formed by both short and long distance connections that help to maintain the Small-World network structure of the graph. To see that, we made an histogram with the distances of all the reliable edges and the result, shown at figure 3.6-a, was as expected.

The graph formed by the reliable edges whose correlation with age has a positive slope has also Small-World values higher than one. However, having a positive slope means that these edges are edges whose weight is strengthened with the age, which is the contrary to what happens with long distance edges. Therefore, we could expect that this graph is formed by short and medium distance connections. The loss of the long distance connections can mean that the graph becomes disconnected, fact that would give us a disconnected graph with separate clusters. This scenario would also maintain a high CC and a low AMPL (that now would measure a within cluster AMPL) values, and would explain the Small-World coefficient value that we obtained. In the distances histogram of this graph (figure 3.6-b) we can see, as expected, that this graph is formed mainly by short and medium distance connections.

Finally, we would expect that the graph formed by the reliable edges whose correlation with age has a negative slope is formed mainly by long distance edges. If that was the case, the graph would maintain a low AMPL value but would lose the high CC value after losing the short distance connections, resulting in a loss of the Small-World structure. In the distances histogram of this graph (figure 3.6-c) we can clearly see that this is the case and the graph is formed mainly by long distance connections.

After seeing the distances distributions of each graph we can conclude that the loss of Small-World network structure of the third scenario was expected, and nothing indicates that this fact can have a negative impact in the results obtained in the previous section.

3.2 Discussion

Prediction of the age of a brain by graph analysis is a problem that can be approached in many different ways. In this article we studied if the edges of the graph separately can work as predictors of the brain age. We found that the weight of some edges (762 out of 4005) have a certain correlation with the age of the brain and some of them are capable of predicting its age with an error of around 8.5 years.

We studied the importance of the negative weighted edges, and we concluded that they do not add relevant information to the graph and thus can be excluded when computing graph descriptors in order to study brain networks. We concluded that positive weighted edges can be good predictors of the edge but, in general, it is the evolution of the weight, and not its sign, which makes an edge to be a good predictor of the age.

We also studied how the graph descriptors Clustering Coefficient, Average Minimum Path Length and Global Efficiency can help us when predicting the age of a brain network. We tested both the binary version and the weighted version of each descriptor, and we found that they both produced very similar results. Therefore, we decided to use the binary versions of the descriptors since are easier to compute. We computed the three descriptors upon three different graphs: a graph formed by all the edges whose weight values through the 97 subjects have certain correlation with their ages, a graph formed only by the subgroup of these edges whose correlation with the age has a positive slope and a graph formed only by the subgroup of these edges whose correlation with the age has a negative slope. We found that the Global Efficiency descriptor in this last group can predict the age with a minimum RMSE value of 7.6 years, which is the best error that we have found throughout all our studies.

In this last scenario, we noticed that the resulting graph lost its Small-World network characterization, but we exposed the reasons why this happens and that this fact does not have an impact in the result of our studies. It is well known that the long distance edges tend to decrease its weight with the edge and, thus, it can be expected that a graph formed only with those edges is a good predictor of the age as a whole. In our experiment we proved this to be certain.

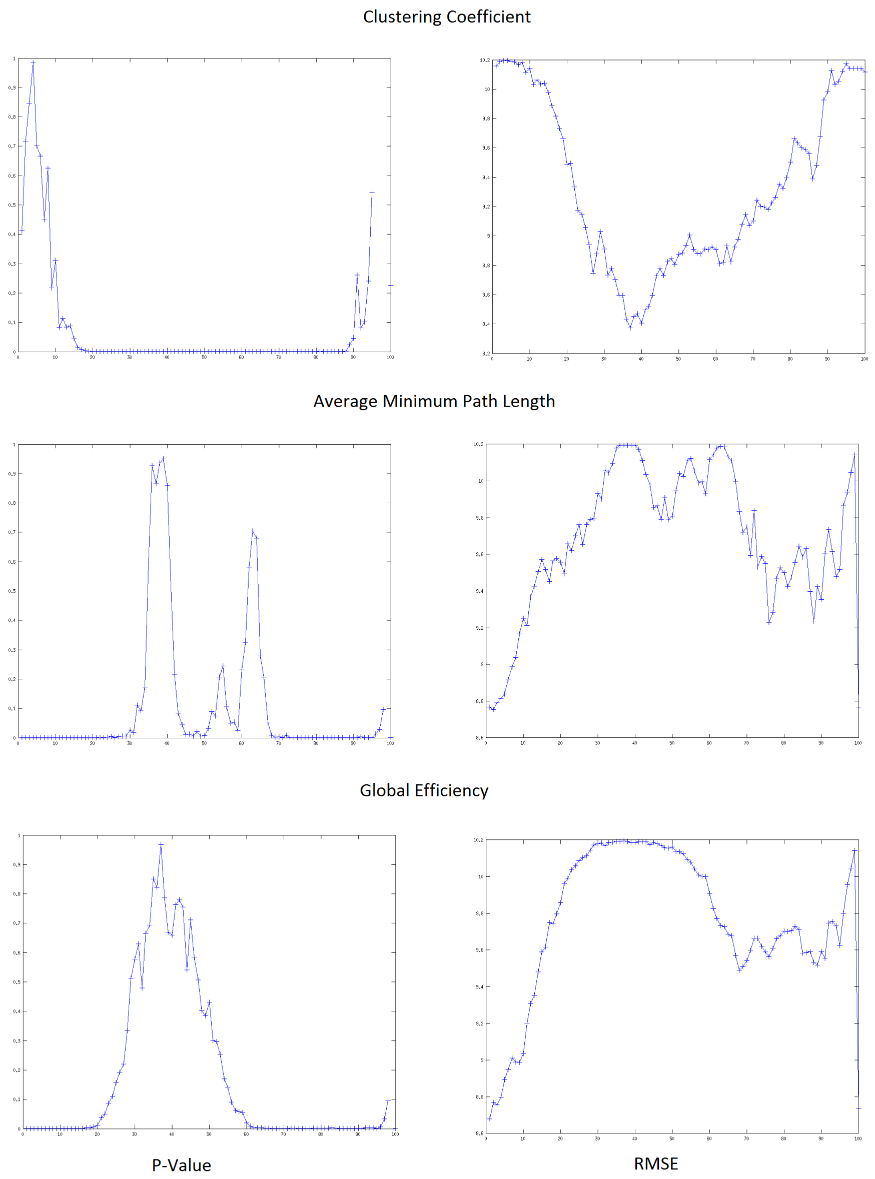


Figure 3.3: Graph formed by all the reliable edges. P-Value and RMSE values of the three descriptors for each of the one hundred different threshold values.

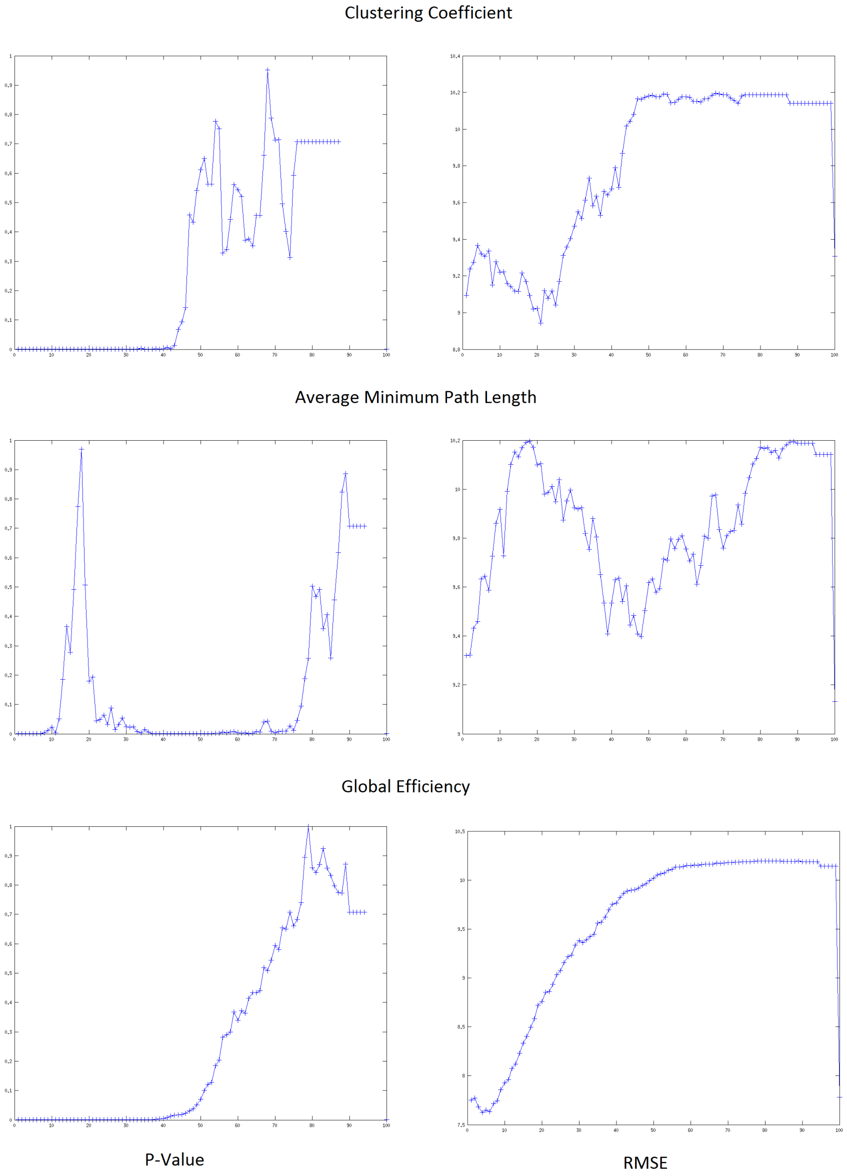


Figure 3.4: Graph formed by the positive reliable edges. P-Value and RMSE values of the three descriptors for each of the one hundred different threshold values.

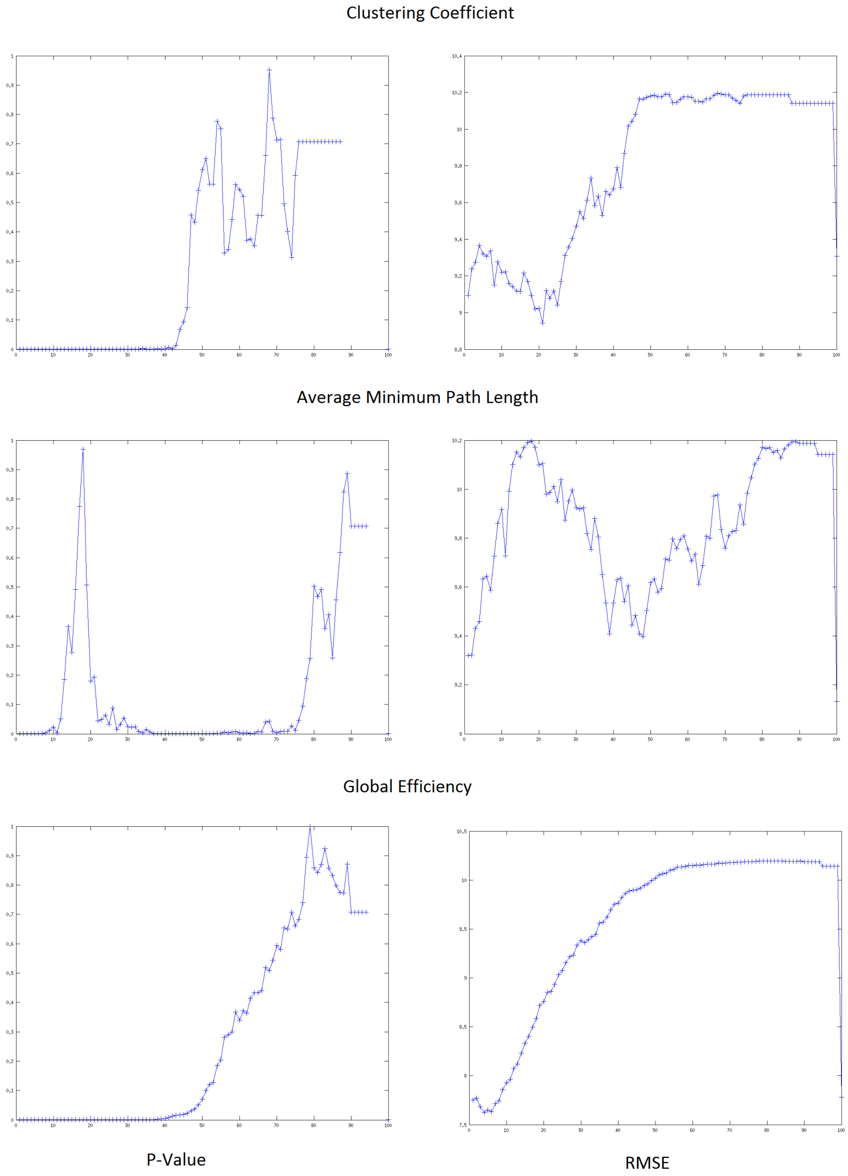
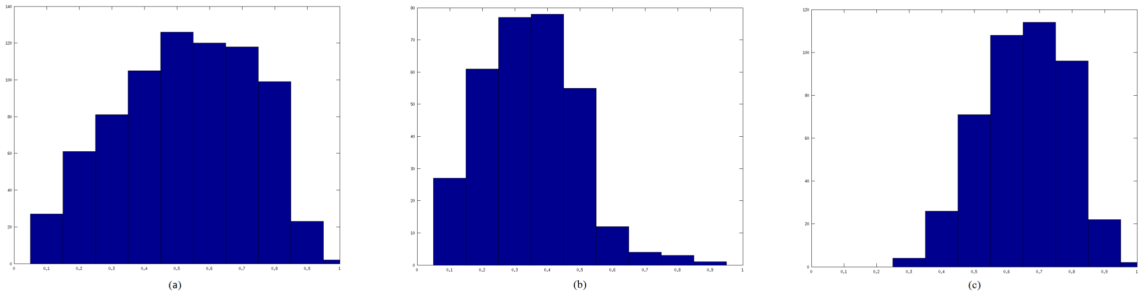


Figure 3.5: Graph formed by the negative reliable edges. P-Value and RMSE values of the three descriptors for each of the one hundred different threshold values.



(a)

(b)

(c)

Figure 3.6: Histograms that show the number of edges classified by its distances.

Chapter 4

Application functionalities

In this project, we developed our own tools in order to compute all the data studied and to obtain the results. More precisely, a total of eight C++ classes were developed and each of them has several methods that we explain in this chapter.

4.1 Main classes

pfcbrain_segmentation.cpp: This class is the core class of the application and its main purpose is computing the segmentation process. Its most important methods are:

- init_pfcbrain_boost: returns an inicializated pfcbrain_boost object. Creates an adjacency matrix from the regions that have been read.
- init_pfcbrain_boost_hierarchy: returns an initialized pfcbrain_boost object from a set of regions that can already contain a certain hierarchy. Its adjacency matrix is created from the regions that have been read.
- createLinks: creates and stores the links between nodes using the specified similarity function and threshold.
- create_link: auxiliary function that creates a new link between the given regions using the specified criteria for comparing their edges, the similarity function and threshold order.
- simple_neighbors: returns true if the specified regions are simple neighbours.
- computeDistance: computes the distance between the two given regions.
- create_bpt: creates the binary partition tree that contains the segmentation of the regions. It stops when the specified number of minimum regions is reached.

- merge_regions: function that generates a new region from a given link that connects two regions.

- compute_region_data: computes the region data for the new region from the data of the merging regions.

- update_links: updates the links BST by creating new links for the new region and deleting links from the old regions.

- find_link: auxiliary function that returns true if the specified link exists in the links BST.

- insert_link_BST: auxiliary function that inserts a new link in the global link BST and in the new_region's link BST.

- erase_link_BST: auxiliary function that removes an existing link of the global links_BST.

pfcbrain_boost.cpp: This class integrates our data structures with the $C++$ Boost library and uses some of its graph analysis tools. It also implements the main descriptors of the graph used in this project. Its most important methods are:

- pfcbrain_boost: creates the hvc_boost object from an existing adjacency matrix.

- BubbleSort: sort algorithm useful for small lists, and for large lists where data is already sorted.

- calculaThreshold: computes the threshold from a given percentage of links that want to kept.

- betweennesCentralityClustering: performs the betweenness centrality clustering upon our data structures.

- clusteringCoefficient: computes the clustering coefficient of the graph.

- weightedClusteringCoefficient: computes the weighted (non-binary) clustering coefficient of the graph.

- shortest_paths: computes the shortest paths of each pair of regions of the graph, and computes and returns the Average Minimum Path Length and Global Efficiency descriptors.

- invertDistancesMatrix: auxiliary function that reverses the distances matrix. Useful for computing some graph descriptors.

- smallWorld: computes the Small-World coefficient of the graph.

4.2 Data Structures

The main data structures used in this application are implemente in three different classes: region, binary partition tree and binary search tree.

- region.cpp: This class represents a region or node of the brain. It also defines the link structure, which represents a link between two regions. For each region, the data structure stores its id, its position in the 3D space (X, Y and Z coordinates), the data of the region, a list of its links, its parent region and both of his son regions. For each link, this structure stores its id, a pointer to the two region which it was formed, and its order.
- b_partition_tree.cpp: This structure represent a concrete partition of the segmentation tree. Their nodes store the region data.
- b_search_tree.cpp: Structure used to store the links tree during the process of segmentation. Very useful due to its speed when doing searches and sorting its elements.

4.3 Utils

- fileParser.cpp: Auxiliary class developed in order to execute all the readwrite from file tasks.
- dijkstra_shortest_path.cpp: Class that implements Dijkstra's single source shortest path algorithm for a graph represented using adjacency matrix representation.
- pfcbrain_utils.cpp: Class with utils. It implements the next functions.
  - same_series: compares two time series and returns true if they are the same.
  - pearson_correlation: given two time series, it computes and returns their Pearson correlation.
  - merge: merges two sorted vectors into one sorted vector.
  - calc_first_derivative: computes the first derivative of the given time series.
  - calc_second_derivative: computes the second derivative of the given time series.

Chapter 5

Working method

This project started officially in April, 2014, when the tutor of the project, Prof. Ferran Marques, and I met the Dr. Roser Sala for the first time. The Dr. Sala has been helping and guiding us through all the project giving her medical perspective, and also is the person that provided all the data necessary to develop this project. From June, 2013 to March, 2014 a previous work related to liver health was started, but due to external issues we had to abandon that project.

During this project, I have been developing my own tools in C++ in order to perform all the experiments that were required. This tools used the data obtained from the fMRI to perform all the experiments described in this document, and provided output files that were later analysed with Matlab in order to create graphs and some other useful metrics that could help us when analysing the data. The results obtained in each experiment were always analysed by the tutor of the project and myself, and then discussed with Dr. Sala in order to give them a medical meaning.

Overall, this project has taken over 2 years to be completed. The main reasons of this were, firstly, that I was working the most part of the time during the project and also started a master course recently. In addition, this kind of projects tend to be slow that it is not always easy to coordinate three people which have their own projects and work in different fields.

Chapter 6

Conclusions

The main objectives of this project have been accomplished. One of the main goals of this project was to improve the age estimation of the human brain and, as proved during this document, we managed to achieved this goal. In the process, we obtained a deep understanding in how the brain-research community works and which are the image processing techniques and research areas that are being studied nowadays in order to improve the understanding of the human brain. Personally, this project has provided to me a great knowledge both in the image processing and brain research areas and have expanded a lot what I learnt during my degree.

A secondary objective of this project was to develop our own tool in order to perform all our experiments. We also managed to accomplish that goal, as we developed some classes and many functions that can be used in order to perform segmentation and graph analysis operations. For me as a student, this part is also very important because this is the first big application that I developed using the C++11 language, and this has been useful to expand the C++ knowledge obtained during my degree. Also, as a Java specialist, this project has helped me to learn another important language which makes me a better programmer overall.

Appendix A

Figures from negative correlations study

All the figures ahead show the computed P-value for the four descriptors used in this study (clustering coefficient, average minimum path length, global efficiency and Small-World coefficient) depending on the chosen threshold used for binarizing the graph in three different scenarios. Figures from A.1 to A.4 show the scenario 1, which keeps the specified percentage of edges with higher values. Figures from A.5 to A.8 show the scenario 2, which keeps the specified percentage of edges with lower values. Both scenarios perform a scrolling of the threshold from 1% to 100% of the edges, with 1% steps. Figures from A.9 to A.12 show the scenario 3, which uses a symmetric threshold using direct edge values and scrolling from 0.01 to 1, with 0.01 steps.

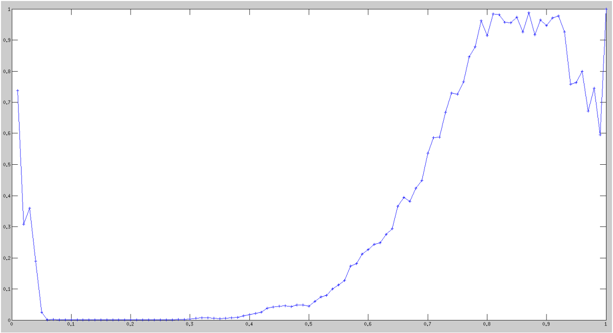


Figure A.1: Higher values. Clustering Coefficient.

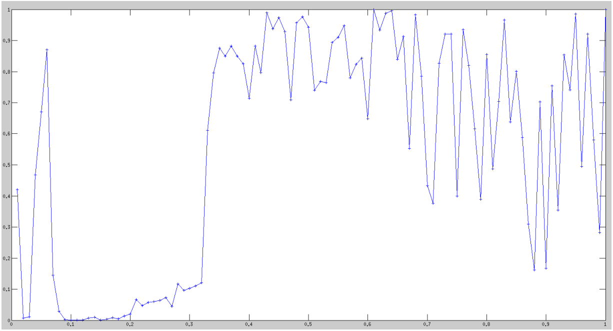


Figure A.2: Higher values. Average Minimum Path Length.

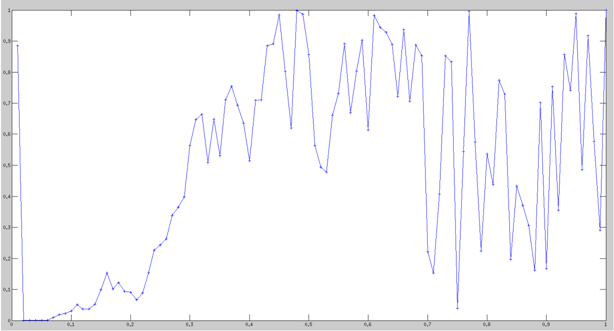


Figure A.3: Higher values. Global Efficiency.

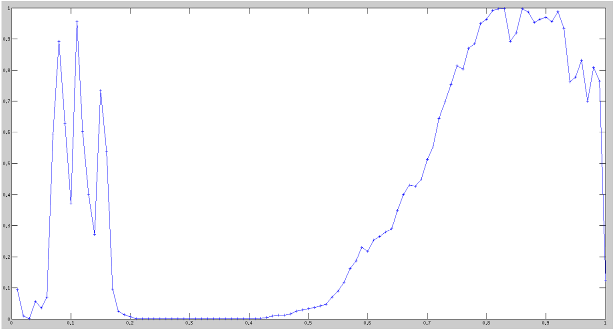


Figure A.4: Higher values. Small World Coefficient.

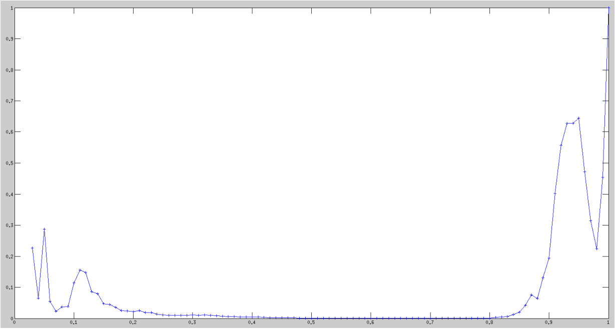


Figure A.5: Lower values. Clustering Coefficient.

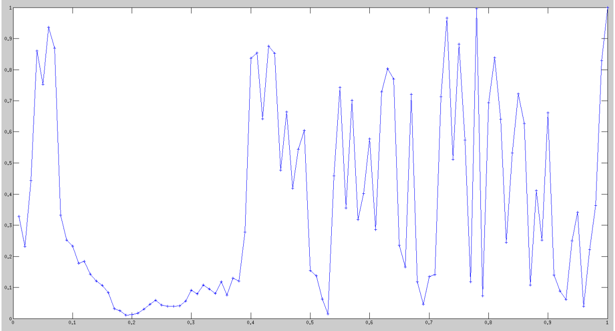


Figure A.6: Lower values. Average Minimum Path Length.

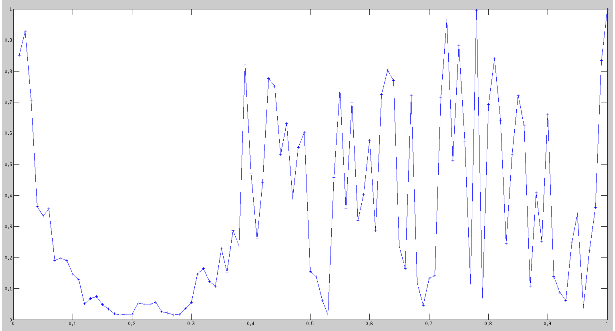


Figure A.7: Lower values. Global Efficiency.

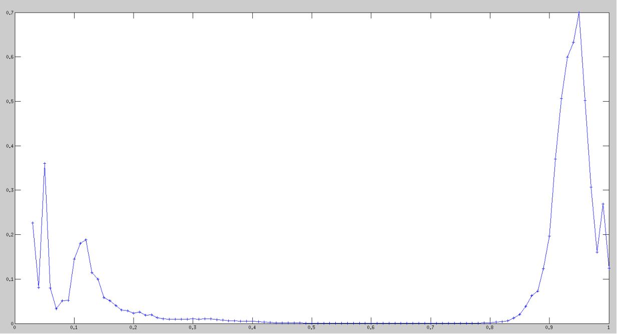


Figure A.8: Lower values. Small World Coefficient.

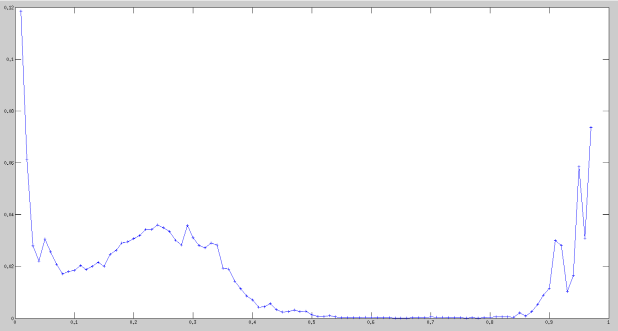


Figure A.9: Symmetrical threshold. Clustering Coefficient.

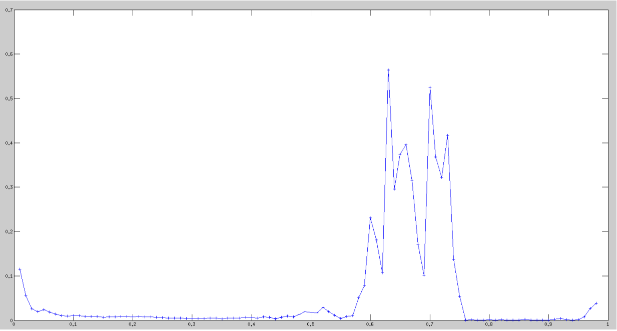


Figure A.10: Symmetrical threshold. Average Minimum Path Length.

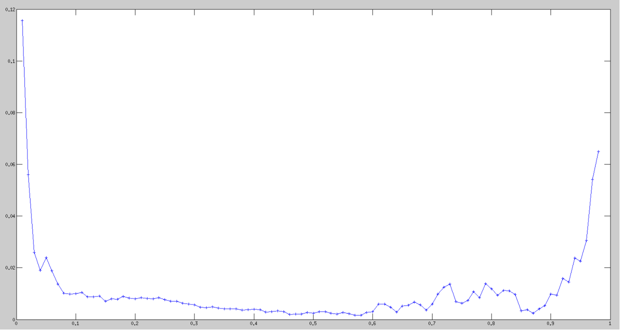


Figure A.11: Symmetrical threshold. Global Efficiency.

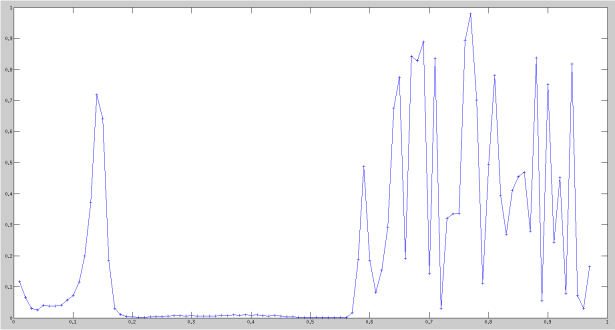


Figure A.12: Symmetrical threshold. Small World Coefficient.

Bibliography

[1] http://www.alzheimers.net/resources/alzheimers-statistics/

[2] http://www.who.int/mental_health/management/schizophrenia/en/

[3] http://www.who.int/mediacentre/factsheets/fs999/en/

[4] https://www.humanbrainproject.eu/es/home

[5] http://www.europeanbraincouncil.org/

[6] Rubinov, M. and Sporns, O., *Complex network measures of brain connectivity: Uses and interpretations*, pp. 1059-1068, October 2009.

[7] Sala-Llonch, R., Junqu, C., Arenaza-Urquijo, E.M., Vidal-Piero, D., Valles-Pedret, C., Palacios, E.M., Domnech, S., Salv, A., Bargall, N., Bartrs-Faz, D. *Changes in whole-brain functional networks and memory performance in aging*, Elsevier Inc., October 2014.

[8] Thompson, P. M., Dennis, E. L., Jahanshad, N., Toga, A. W., McMahon, K. L., Zubicaray, G. I., Martin, N. G., Wright *Test-Retest Reliability of Graph Theory Measures of Structural Brain Connectivity*, pp. 305-312, 2012.

[9] Meier, T.B., Desphande A.S., Vergun, S., Nair, V.A., Song, J., Biswal, B.B., Meyerand, M.E., Birn, R.M., Prabhakaran, V., *Support vector machine classification and characterization of age-related reorganization of functional brain networks*, Elsevier Inc., March 2012.

[10] Meunier, D., Achard, S., Morcom, A., Bullmore, E., *Age-related changes in modular organization of human brain functional networks*, February 2009.

[11] Alexander-Bloch, A.F., Gogtay, N., Meunier, D., Birn, R., Clasen, L., Lalonde, F., Lenroot, R., Giedd, J., Bullmore, E.T., *Disrupted modularity and local connectivity of brain functional networks in childhood-onset schizophrenia*, October 2010.

[12] Bassett, D.S., Bullmore, E., *Small-world brain networks*, December 2006.

[13] Latora, V., Marchiori, M., *Efficient Behavior of Small-World Networks*, February 2008.

[14] Smith, S. L., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D., Woolrich, M. W. *Network modelling methods for fMRI*, pp. 875-891, 2011.