Validating the All-Resolutions Inference method for analyzing fMRI and building a comprehensive app for the end users

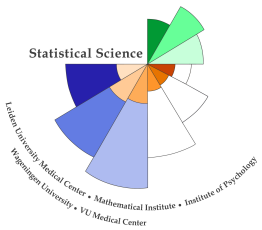
Martha van Kempen (s1972723)

Thesis advisor: Dr. Wouter W. Weeda
Second thesis advisor: Prof. Dr. Jelle J. Goeman

MASTER THESIS

Defended on March 29, 2019

Specialization: Data Science



STATISTICAL SCIENCE
FOR THE LIFE AND BEHAVIOURAL SCIENCES

Contents

Chapter 1

Introduction

1.1 The use of fMRI

Functional magnetic resonance imaging (fMRI) is a widely used method to localize brain activity in response to cognitive stimuli in patients and healthy subjects. The blood flow inside the brain is measured during an fMRI scan by looking at the ratio of oxygenated and deoxygenated hemoglobin. This ratio is called the blood oxygenation level-dependent (BOLD) effect. When a brain region becomes active, more oxygen is needed in that area which leads to an increase in blood flow towards the active brain area. This change in oxygen levels in response to cognitive stimuli is measured by fMRI to locate the brain activity. The first article on PubMed about fMRI was published in 1993 by Bandettini et al. (1993). Over the past 25 years, acquisition and post-processing methods of fMRI have greatly improved. Nowadays, there are approximately 40000 articles on PubMed that mention fMRI in their title or abstract. Despite the tremendous amount of research done, the interpretation of fMRI scans remains challenging. The next section will explain this challenge in the most used fMRI analysis method.

1.2 Current most used analysis method in fMRI research

One of the reasons that the interpretation of fMRI scans remains challenging is that the statistical analysis of the scans is difficult and not yet optimized. A scan of the brain consists of possibly millions of voxels for which activity has to be calculated. Performing many tests is called multiple testing which is a statistical problem that can lead to false positives when no correction method is used. The Bonferroni is a correction method for multiple testing problems, but this results in very low power. Therefore, the Bonferroni method is not the best way to find activity in fMRI scans. Cluster-extent based thresholding is another method that is suited for multiple testing problems and has a higher power than the Bonferroni method. Therefore, the most widely used analysis method in fMRI studies and software programs for locating brain activity is cluster-extent based thresholding. This method divides the brain into clusters of voxels by removing the voxels that are below a primary threshold. After that, the cluster-level extent-threshold is determined which controls the probability of finding false positives (Woo et al., 2014). The cluster level extent-threshold represents the number of contiguous voxels and is often estimated by random field theory (RFT). This method has high sensitivity, but Woo et al. (2014) found that the cluster-extent based method has low spatial specificity. Spatial specificity is very important in making interpretations about brain areas that are activated by a cognitive task. When a cluster is found active by the cluster-extent based thresholding, not all voxels in that activated cluster are active, but it merely means that at least one voxel within that cluster is active. The clusters

in cluster-extent based thresholding are often very big and cover multiple brain regions which makes it hard to identify the specific location of the brain activity. The brain has much more voxels than it has brain regions which shows that having one active voxel is not representative for multiple brain regions. This makes it difficult to make conclusions about which brain regions are really affected by the performed task during the fMRI scan. The bigger a cluster is, the harder it is to localize the activity. This is called the spatial specificity paradox. For this reason, the ability to locate brain activity more precise is important for future improvements in analysis methods. Luckily, a lot of research is done to find a better analysis method. Thereby, psychology is not the only field that has to deal with the multiple testing problem. One of the biggest fields that work with the multiple testing problem is genomics. Analysis methods that are designed for the genomics field can be translated to fMRI analysis. The next section will discuss a newly developed analysis method that is better in localizing brain activity and is of more exploratory nature than the cluster-extent based thresholding method.

1.3 All-Resolutions Inference as new method

A big improvement on the cluster-extent based thresholding method which was discussed in the previous paragraph would be the ability to drill down from a big cluster of activation to smaller sub-regions of activation within an fMRI scan. This improves the local specificity because the clusters become smaller and there are fewer voxels in which the activity could be located. However, making selections after performing the analysis is often statistically invalid because it can lead to optimistic results caused by dropping voxel-wise error guarantees. Rosenblatt et al. (2017) showed that the All-Resolutions Inference (ARI) method of Goeman and Solari (2011) do give valid results with a comparable problem. It was shown that ARI has the ability to drill down within clusters that were formed based on statistics values of the same fMRI scan while holding the error guarantees (Rosenblatt et al., 2017). This is one of the biggest advantages of ARI compared to other analysis methods because it allows a more exploratory approach to finding activity within an fMRI scan. The ARI method is based on the closed testing procedure of Marcus, Peritz, and Gabriel (1976) and uses local Simes tests and a shortcut provided by Goeman et al. (2017) to control the FWER. The Simes inequality has to hold for the Simes test to be valid. Luckily, a lot of research has been done to the Simes inequality since this assumption is also necessary for multiple other methods like Benjamini and Hochberg (1995). Nichols and Hayasaka (2003) imply that the Simes inequality holds with the positive regression dependency on subsets condition (PRDS; Benjamini and Yekutieli, 2001) for brain maps. Therefore, the assumption that the ARI method is suited for analyzing brain maps is made. The ARI method can calculate the proportion of true discoveries (PTD) for all possible clusters the researcher is interested in that are based on the same data. The PTD describes the proportion of active voxels in a cluster. Some clusters have a high PTD, while other clusters have a low PTD depending on the data and threshold settings. The researcher is allowed to use multiple thresholds, change settings during and after the analysis and use multiple drill down approaches after the first analysis. Having a PTD for small clusters is a big improvement for the local specificity paradox. For example, if the ARI method returns a big cluster with 70% of active voxels, the researcher is allowed to increase the threshold of a cluster which could lead to a smaller cluster with a higher PTD. After that, the same can be done for other clusters or the smaller new clusters. The threshold can be changed for multiple clusters, but also for brain regions, spheres and local optima in the image. This can all be done while maintaining proper FWER control. This shows that settings can be changed after doing the analysis and thresholds can be changed multiple times which makes the ARI method very flexible and suitable for exploratory research which is one of the main advantages of the ARI method. This in combination with dividing big clusters into smaller clusters repeatedly, makes

the interpretation of the location of the activity easier, more precise and more reliable than the cluster-extent based thresholding method. The theoretical background of the ARI method will be explained in chapter 2 in which the explanation and the structure of the article of Goeman and Solari (2011) is used as the foundation. The next section will describe outstanding questions about the ARI method that still need to be answered.

1.4 Outstanding questions

The All-Resolutions Inference (ARI) method was proven to be effective on both fMRI simulation data and real fMRI data (Rosenblatt et al., 2017), but there are still many questions that need to be answered. The main question of this report is:
Is the All-Resolutions Inference method able to give valid results about activity in fMRI scans and can researchers easily apply the ARI method?
This question can be divided into three parts and each part will be discussed in a separate chapter. Testing the ARI method on null data is a good next step in validating the ARI method since previous research (Eklund et al., 2016) showed that several software programs for analyzing fMRI scans gave false positives when applying task designs on null data.

1. This leads to the first question: What is the performance of the All-Resolutions Inference (ARI) method on fMRI scans unexposed to stimuli in comparison to the most widely used fMRI software program FSL? The hypothesis is that the ARI method performs better on null data than FSL and that the false positive rate of ARI will not exceed the 5%.

The previous question will investigate the false positive rate of the ARI method, but the false negative rate is also important. The false negative rate describes the proportion of voxels where the ARI method finds no activity when there actually is an activity. The ARI method should be able to find activity in fMRI scans and the power of the method is for this reason important. The power is defined by subtracting the number of false negatives (FN) of one ($1 - FN$). The power can be calculated by comparing the ARI results with the results of other methods performed on the same fMRI data.

2. The second part will answer the question: How does the All-Resolutions Inference method perform in identifying activity on fMRI scans in comparison to previously described methods in literature? The hypothesis is that the ARI method will have sufficient power to find activity in fMRI scans.

One disadvantage of ARI is that it can be hard to perform without prior knowledge of the use of the method. The method has many different options and it can be difficult to perform all those options or decide which option is best. Therefore, a guide for the user would be very helpful. Besides that, a visualization of the results can help the user to interpret the results. Since the ARI analysis is very suitable for exploratory analysis, an interactive app to guide the user will be most useful. There is a specific package developed in R to perform the ARI method called ARIbrain (Rosenblatt et al., 2017). Therefore, building an app within R is most suited and the package shiny can be used (Chang et al., 2018).

3. The third part will discuss the building and use of an interactive Shiny app in R to support the end-user in applying the ARI method during the analysis of fMRI scans. All the different options can be tried out within the app which will help the researcher in the analysis of fMRI scans. The app will be accompanied by an user manual.

1.5 Outline and approach

The structure of this paper follows the order of the questions mentioned above. Chapter 2 will discuss the theoretical background of the ARI method to give an understanding of the method to the readers. Chapter 3 will discuss the first part of the research question about the performance of the ARI method on null data. For this purpose, a part of the data that was used by Eklund et al. (2016) will be used. These data will also follow the same pre-processing settings in FSL as was used by Eklund et al. (2016). The fourth chapter will discuss the second part of the research question about investigating the power of the ARI method. All available data of the Neurovault database were downloaded, but many scans were useless and therefore filtered out. The ARI method was applied on the remaining scans and the ARI results were compared to the results discussed in the articles accompanying the scans. After validating the ARI method, the fifth chapter will discuss the building of the interactive app and give step by step instructions on how to use the app. A real example will be used in the tutorial of the app.

Chapter 2

Theoretical Background ARI method

This chapter uses the structure of the article of Goeman and Solari (2011) as a foundation. This includes the examples and background knowledge discussed in the article.

2.1 Introduction

Most multiple comparison correction methods for fMRI scans are based on the family-wise error rate (FWER) or the false discovery rate (FDR). These are both analysis methods that can deal with multiple testing problems. The probability of making any false rejection is controlled by FWER based methods at a prespecified rate. When this is translated to the analysis of fMRI scans, these methods control the probability of finding any false activity in a voxel in the brain. An example of an FWER based method is the Bonferroni method which has a strong FWER control. One problem of the FWER based methods is the vanishing power as the number of hypotheses increases. The other procedure on which many multiple comparison correction methods are based is called FDR. FDR-based methods control the expected proportion of false discoveries among all discoveries found. FDR based methods have weak FWER control. An example of an FDR based method is the BH procedure (Benjamini and Hochberg, 1995). The power of FDR based methods does not vanish when there is enough signal and remains stable when the number of hypotheses increases (Chi, 2007). FWER and FDR based methods are types of error control that are often used in hypothesis testing during confirmatory analysis. A confirmatory analysis is done to confirm a specific hypothesis that is based on previous research. When there are not yet any hypotheses formed about the data, an exploratory analysis is used to discover trends and potential hypotheses for further research. FMRI analysis is more suited for exploratory research since a researcher wants to find activity in an image that exists of possibly millions of voxels without a specific hypothesis for each voxel. Therefore, a more exploratory analysis than FWER and FDR is needed. This chapter will describe an exploratory analysis that is suited for fMRI analysis. An analysis is exploratory when it fulfills the three conditions of being mild, flexible and post-hoc (Goeman and Solari, 2011). A mild analysis tolerates some false positives among the rejected hypotheses. False positives that are returned during exploratory research will be removed during the follow-up research. Confirmatory research should exclude false positives immediately from their analysis and needs a strict analysis because there will be no follow up after confirmatory analysis. However, it is very time-consuming and exhaustive if all rejected hypotheses during exploratory analysis have to be followed up on. To avoid following up on hundreds of false positives, some multiple testing error corrections have to be applied. A good trade-off between the number of false positives and false negatives has to be found.

A flexible approach should allow the researcher to select rejected hypotheses to follow up on and not force the researcher to pick one. The researcher may find a connection between rejected hypotheses which can only be noticed through experience which the analysis itself cannot do. The last condition of exploratory research is being a post-hoc procedure. This means that all decisions about the analysis can be made after analyzing or after gathering the data. The data and results can give the researcher information on how mild and flexible the analysis should be. All FWER based methods make a decision about the control rate prior to the analysis. Therefore, this collection of methods is not post-hoc and not mild. The FDR based methods are milder compared to the FWER based methods. However, FDR based methods do set the FDR threshold before the analysis which makes this group of methods not post-hoc as well. Furthermore, flexibility is also missing in the FDR based methods. Only subsets of the rejected hypotheses can be chosen that stay below the pre-specified FDR threshold. This means that both the FDR based methods and the FWER based methods do not meet all three conditions. This shows the importance of the development of an exploratory analysis that is mild, flexible and post-hoc which will be discussed in the next sections.

2.2 Closed testing procedure

All-Resolutions Inference (ARI) is an exploratory analysis method that fulfills all three conditions mentioned in the previous section. ARI is based on the closed testing procedure of Marcus, Peritz, and Gabriel (1976). This method is combined with the confidence set and the discrete version of the confidence interval (Goeman and Solari, 2011). A confidence set includes the true parameter in $(1-\alpha)*100$ percent of the cases, in which α represents the confidence level. This combination of procedures is able to give confidence bounds on the number of false rejections when a specific set of hypotheses is rejected. The closed testing procedure is based on the collection of all hypotheses that are tested namely the elementary hypotheses, and the intersection hypotheses which are the intersections of elementary hypotheses. In fMRI research, each voxel has an elementary hypothesis which can possibly result in millions of elementary hypotheses. Within the collection of all elementary hypotheses, some are false and some are true. The indices that represent true hypotheses are stored in $\mathcal{T}$. The indices of the true hypotheses are not known beforehand. Intersection hypotheses can be made from the elementary hypotheses and are called H_I. I represents the subset of indices of the elementary hypotheses that are used in that intersection hypothesis. All true and false intersection hypotheses are stored in $\mathcal{C}$. An intersection hypothesis is true when all elementary hypotheses with indices in I are also true. The closed testing procedure has to perform a local significance test on all intersection hypotheses. The rejected hypotheses uncorrected for multiple testing are stored in $\mathcal{U}$. Every intersection hypothesis is rejected when at least one index of the intersection hypothesis is stored in $\mathcal{U}$. This closed testing procedure controls the FWER for all intersection hypotheses (Marcus, Peritz and Gabriel, 1976).

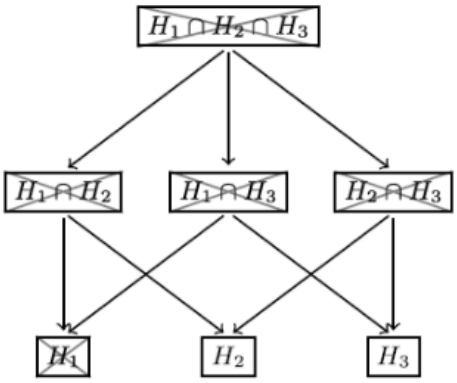


Figure 2.1: Results of the closed testing procedure on intersection hypotheses formed by elementary hypotheses. Image from Goeman and Solari (2011).

A consonant procedure results in higher power compared to non-consonant procedures. Consonant means that when an intersection hypothesis is rejected, at least one underlying hypothesis is rejected as well. Figure 2.1 from Goeman and Solari (2011) shows a consonant rejection for the intersection hypotheses $H_1 \cap H_2$ and $H_1 \cap H_3$. A non-consonant rejection means that the intersection hypothesis is rejected while none of the underlying hypotheses are rejected. The intersection hypothesis $H_2 \cap H_3$ in figure 2.1 represents a non-consonant rejection. Consonance leads to a higher power in FWER based methods and is therefore often desirable. However, procedures that do not allow non-consonant rejections are more strict compared to a procedure in which non-consonant rejections are allowed. Mild procedures are better for more exploratory analysis and therefore, allowing non-consonant rejections improves the analysis of fMRI. The procedure of Hommel (1988) is an FWER controlling method that uses the Simes test which does allow non-consonant rejections. Therefore, the Simes test is also used in the ARI method to make the method mild. The Simes test is explained in the next section.

2.3 Simes

The computation time of fMRI analysis will increase when the number of elementary hypotheses increases. An fMRI scan exists of many voxels which means that the computation time would be very large if there were no shortcuts available. Shortcuts provide a reduction of computations that have to be made to get the same or approximately the same solution. A shortcut for the closed testing procedure can reduce the number of local tests that have to be performed which will reduce the computation time. The All-Resolutions Inference (ARI) method uses the Simes test as local test. This local test rejects an intersection hypothesis when at least one p-value of the elementary hypotheses is below a specific critical value.

$$p_{(i:I)} \leq c_{(i:I)} \tag{2.1}$$

$p_{i:I}$ stands for the p-values of the elementary hypotheses from which the intersection hypothesis is made of. $c_{i:I}$ represents the critical values accompanying the p-values of the elementary hypotheses from the intersection hypothesis. The p-values and critical values are sorted in ascending order. There are as many critical values as there are hypotheses that are tested by the closed testing procedure.

The critical values are calculated with this formula:

$$c_{(i:I)} = \frac{i\alpha}{|I|} \tag{2.2}$$

Where α is the significance level and $|I|$ is the number of indices in the intersection hypothesis. The combination of these two formulas leads to the rejection of an intersection hypothesis when:

$$|I|p_{(i:I)} \leq i\alpha \tag{2.3}$$

for at least one value of $1 \leq i \leq |I|$. However, the Simes test is only valid when the Simes inequality holds for true hypotheses. The Simes inequality says that the probability of rejecting at least one true intersection hypothesis should be below α. The Simes inequality holds for independent p-values and for positive correlations of certain forms. Since a hypothesis is rejected when equation (2.3) is true, the combination of the Simes inequality and equation (2.3) returns the following formula:

$$P(|T|p_{(i:|T|)} \leq i\alpha) \leq \alpha \tag{2.4}$$

Since there are more analysis methods that use the Simes test like Hommel (1988) and Benjamini and Hochberg (1995), a lot of research has been done on the Simes inequality (Benjamini and Yekutieli, 2001; Rødland, 2006; Sarkar, 2008; Finner et al., 2014). Nichols and Hayasaka (2003) imply that the Simes inequality holds for brain maps with the positive regression dependency on subsets condition (PRDS; Benjamini and Yekutieli, 2001). Therefore, the assumption is made that the Simes Inequality holds and that the Simes test is valid for the ARI method. The Simes test can now be used in the calculation of the proportion of true discoveries (PTD) which will be discussed in the next section.

2.4 Calculating the proportion of true discoveries

This section will discuss the calculation of two confidence sets with the help of the All-Resolutions Inference (ARI) method. A confidence set can be calculated for the number of false rejections and also for the number of false hypotheses within a set of hypotheses. These confidence sets can also be calculated for the schedule in figure 2.1. The maximum number of false rejections for one hypothesis is based on the first ascendant hypothesis that is not rejected. The $H_I = H_1 \cap H_2 \cap H_3$ is an intersection of three hypotheses. The three ascendant hypotheses from this H_I are each the intersections of two elementary hypotheses: $H_1 \cap H_2$, $H_1 \cap H_3$, $H_2 \cap H_3$ as shown in figure 2.1. Those three ascendant intersection hypotheses are all rejected by the closed testing procedure. If you look again in figure 2.1, there are two elementary hypotheses H_2 and H_3 that are not rejected. These two hypotheses are the first ascendants that are not rejected and they exist of only one index number since they are elementary. This means that the maximum value in the confidence set for the number of false rejections is one. This value is also called the $t_\alpha(S)$. The confidence set for false rejections always includes the values between zero and the $t_\alpha(S)$. The false discovery proportion (FDP; $q_\alpha(S)$) is the $t_\alpha(S)$ divided by the size of the set S. Goeman et al. (2017) proved that the results of the FDP confidence bounds (q_α) do not vanish as the number of hypotheses go towards infinity. The proportion of true discoveries (PTD) is calculated by $1 - FDP$.

To calculate the confidence set of false hypotheses for a H_I which is called $\phi(S)$, the number of indices representing the elementary hypotheses included in the intersection hypothesis have to be counted. This number is the upper bound of the confidence set. The lower bound is the number of elementary hypotheses covered in the intersection hypothesis minus the number of

elementary hypotheses that are part of the first not rejected H_I ($t_\alpha(S)$) used to calculate the previously described confidence set. Both the confidence sets are calculated for figure 2.1 and are displayed in table 2.1.

S	**Confidence set for** $t(S)$	**Confidence set for** $\phi(S)$
{1}	{0}	{1}
{2}	{0,1}	{0,1}
{3}	{0,1}	{0,1}
{1,2}	{0,1}	{1,2}
{1,3}	{0,1}	{1,2}
{2,3}	{0,1}	{1,2}
{1,2,3}	{0,1}	{2,3}

Table 2.1: Defining the confidence sets for the number of false rejections ($t(S)$) and for the false hypotheses ($\phi(S)$) among the indices in S.

The resulting confidence sets can help the user decide which sets of hypotheses to reject. There are many different options and the user has the ability to pick the set of hypotheses to follow up on themselves. This makes this approach flexible. The user can also decide how many false rejections are allowed which makes this procedure mild. And the last condition of an exploratory research method is also met. The analysis is post-hoc since the user can compare the consequences of all the possible sets of hypotheses after the analysis is done. Therefore, the All-Resolutions Inference (ARI) method is very suitable for exploratory research and fMRI research.

2.5 Shortcuts

Besides the fact that the ARI method is named suitable for fMRI research in the previous section, the closed testing procedure performs 2^m local tests which results in a very long computation time when there are many elementary hypotheses. The m stands for the amount of elementary hypotheses. To reduce the number of local tests, a shortcut can be applied to calculate $t_\alpha(S)$. As earlier mentioned, Shortcuts provide a reduction of computations that have to be made to get the same or approximately the same solution. Goeman et al. (2017) developed a new shortcut for the closed testing procedure that is applicable for all S and more exact and faster to compute than the shortcuts mentioned in Goeman and Solari (2011). This new shortcut is used in the ARI method to find activity in fMRI scans. In this shortcut, a set of hypotheses is made consisting of the hypotheses with the largest p-values of size i and this set is stored in K_i with $0 \leq i \leq m$. A set K_i is made for each possible value of i. This K_i is the worst possible set of hypotheses of size i with the highest probability of being accepted. From those worst possible sets of hypotheses, some are rejected and some are not rejected by the local test. Among the K_i's that are not rejected, the K_i that consists of the most hypotheses is called h. If the size of any intersection hypothesis is bigger than h, that intersection hypothesis is rejected by the closed testing procedure. This shows that $t_\alpha(\{1, ..., m\}) = h$. The $(1-\alpha)$-upper confidence bound for the proportion of true null hypotheses ($\hat{\pi}_\alpha$) is therefore $\frac{h_\alpha}{m}$. Another functionality of h is the determination of rejecting or accepting an intersection hypothesis. The intersection hypothesis gets rejected if there exists a value i between 1 and the size of the intersection hypothesis such that $hp_{(i:I)} \leq i\alpha$. The actual shortcut is defined by Goeman et al. (2017) as follows:

$$d_\alpha(S) = \max_{1 \leq \upsilon \leq |S|} 1 - \upsilon + |\{i \in S : hp_i \leq \upsilon\alpha\}|. \tag{2.5}$$

where the maximum can be obtained by an υ between 1 and $|S|$. The υ can also be defined in another way to reduce the computation time. This definition says that the maximum for equation 2.5 can be obtained for an υ at most z_α which is defined in equation 2.6. The time that is needed to calculate z_α is O(m).

$$z_\alpha = \begin{cases} 0 & \text{if } h = m \\ min\{m - h \leq i \leq m : hp_{(i)} \leq (i - m + h + 1)\alpha\} & \text{otherwise} \end{cases} \tag{2.6}$$

The h in equation 2.5 only needs to be calculated once for each α since it does not depend on the selected set S. It takes O($m$log$m$) time to determine the value of h for all α simultaneously. The following formula will provide another reduction of the computation time of equation 2.5.

$$b_{(i:S)} = \frac{hp_{(i:S)}}{\alpha} \tag{2.7}$$

This equation is calculated for all the elementary hypotheses covered in the set S. The outcome values are rounded to above and the values that are bigger than the size of S are removed. The remaining values are sorted in linear time and only those are used to calculate the maximum in equation 2.5. This shows that the actual p-values are not important, but only the amount of multiples they are of the ratio α/h.
The shortcut described in this section is used to calculate the confidence bounds of the FDP.

2.6 Translating All-Resolutions Inference to the analysis of fMRI images

The previous sections show that ARI is an analysis method that is well suited for an exploratory analysis when there are many elementary hypotheses. This indicates that ARI is a good method for finding activity in fMRI images. Each voxel in an fMRI image represents an elementary hypothesis and a cluster represents an intersection hypothesis. The confidence sets for the number of false discoveries ($t_\alpha(S)$) can describe the number of inactive voxels within a cluster. The false discovery proportion (FDP) in the brain is calculated by dividing the lower bound of this confidence set by the total number of voxels in the set. The proportion of true discoveries (TDP) is then calculated by $1 - FDP$. This can be done for all possible sets of voxels. The main advantage of the ARI method described in this chapter is the exploratory nature. Researchers are able to run the ARI analysis multiple times with different clusters of interest. The clusters can be based on the values of a statistical map or be based on brain areas. A threshold has to be set by the researcher before the analysis, but it can be changed after running the analysis. The ARI method also provides the possibility to drill down within clusters that are based on the same fMRI data. Researchers should use the ARI method to get a better estimation of the location of the activity. If big clusters have a low PTD, the ARI method can be used to set a higher threshold for a big cluster to get better local specificity. The PTD will increase for smaller clusters and the interpretation of the activity is easier. Researchers should use all the possibilities that the ARI method provides to get the best analysis of activity ins fMRI data.

Chapter 3

The negative predictive value of the All-Resolutions Inference method on fMRI scans unexposed to stimuli.

3.1 Introduction

It was already mentioned in the introduction of this report that the All-Resolutions Inference (ARI) analysis method was proven to be effective on both task-based fMRI simulation data and real task-based fMRI data (Rosenblatt et al, 2017). This shows that the ARI method can find true positives, but it does not give information about the false positive rate. Previous research has found that multiple software programs that are used for analysis of fMRI scans, return false positives in the cluster-wise inference of brain activity (Eklund et al., 2016). These software programs depend on Gaussian random field theory (RFT) which has the assumption that the spatial autocorrelation function has a squared exponential shape. The results of Eklund et al. (2016) indicate that the cause of this high false positive rate is the heavy-tail spatial autocorrelation which means that the previously mentioned assumption is not met. It is called a false positive when the analysis finds activity in a brain area, while there is actually no activity at all. The software programs that were used in the previous research include SPM, FSL, and AFNI which are all three widely used in fMRI research. The data included resting state fMRI data with no signal from Beijing, Cambridge and Oulu from the 1.000 Functional Connectomes Project (Biswal et al., 2010). Eklund et al. (2016) chose these three data sets because of the narrow age ranges of the subjects and large sample sizes. The 1,000 Functional Connectomes Project data set is nowadays often used to validate techniques and control for false positives in fMRI research. A good next step in the validation of the ARI method is, therefore, to apply the method to null data. The question that arises is, therefore: What is the performance of the All-Resolutions Inference (ARI) method on fMRI scans unexposed to stimuli in comparison to the performance of FSL? The hypothesis is that the ARI method will perform better than FSL, which means that we expect ARI to find less significant brain activation in null data compared to FSl. The expectation is to find a false positive rate of around 5% for the ARI analysis. To answer the question, part of the 1.000 Functional Connectomes Project will also be used in this study. However, only the Oulu data will be used and not the Beijing and Cambridge data because the Oulu data set exists of 103 subjects which will be sufficient for this analysis. FSL is chosen as software program since it is the most widely used analysis method in the fMRI research field. This chapter will describe the experiment in which we compared the FSL and the ARI method.

3.2 Materials and Methods

Descriptives

The data set from Oulu that was part of the 1000 Functional Connectomes Project was used existing of 103 subjects. There were 37 male subjects and 66 female subjects with an age ranging from 20 till 23. The scans were made with a 1.5 Tesla scanner and a repetition time of 1.8 seconds was used. There were 28 slices made and 245 time points per subject. There is a T1 weighted scan available for each subject which can be used for registration.

Preprocessing

The fMRI scans were preprocessed before running the ARI method. To control for the effect of the preprocessing settings on the outcome of the ARI, the data was preprocessed multiple times with different settings. The same FSL pipeline was used that was discussed by Eklund et al. (2016), namely FSL FLAME1. The analysis was done on scans with different values for multiple settings which are displayed in table 3.1. One of these settings is the full width at maximum (FWHM) values for spatial smoothing. The higher the FWHM values, the more smoothing is applied. As you can see in table 3.1, the FWHM values 4 mm, 6 mm, 8 mm and 10 mm are used. The other settings include options of two block designs (B1 and B2), two event-related designs (E1 and E2), two threshold levels and two analysis levels. During the cluster forming, only the z-values above the threshold are used. The cluster-forming thresholds are arbitrary and not fixed on the data. This does not influence the FWER control in the ARI method. The z-threshold of 2.3 is the default option in FSL and the z-threshold of 3.1 is the default option in SPM. SPM is another widely used software program for the analysis of fMRI scans.

Setting	**Values**
Smoothing	4mm, 6mm, 8mm and 10mm
Block Design	B1 (10-s on off) and B2 (30-s on off)
Event Design	E1 (2-s activation, 6-s rest) and E2 (randomized, 1- to 4-s activation, 3- to 6-s rest)
Z Cluster Threshold	2.3 and 3.1
Group size	20 and 40
Analysis Level	2 groups and 1 group

Table 3.1: Values of different preprocessing settings used in FSL including smoothing levels, designs, thresholds, number of subjects within the group and the analysis level.

Analysis

Subsets of subjects were randomly selected to perform one sample t-tests to test for group activation and two-sample t-tests to test for group differences. The groups of 20 subjects or 40 subjects were selected by randomly permuting the subject numbers and selecting the first 20 or 40 of this permuted data set to put in group one and the next 20 or 40 to put in group two. This random permutation of groups was done 100 times and the same 100 permutations were used for each combination of settings. All the combinations of settings and random permutations led to a total of 28800 preprocessed scans. After the preprocessing of the data, both the FSL and the ARI analysis were performed. The ARI analysis was done in R with the package ARIbrain (Rosenblatt

et al., 2017). The Z-maps and masks that were returned by FSL after the preprocessing was used for the ARI analysis as well. Both statistical maps and maps of p-values are needed for the ARI analysis, so the z-maps were converted to p-maps with the help of fslmaths in FSL. The whole mask was used as a cluster map in the ARI analysis, which means that no threshold was used resulting in only one cluster consisting of the whole brain. The cluster forming thresholds were used for the FSL analysis. Significant clusters for each preprocessed scan were calculated for both the ARI method and the FSL method. The value one was assigned to a scan when at least one significant active voxel was found. The scan got the value zero when there were no significant active voxels found. The results of the ARI method and the FSL method were then compared to each other and are discussed in the next section.

3.3 Results

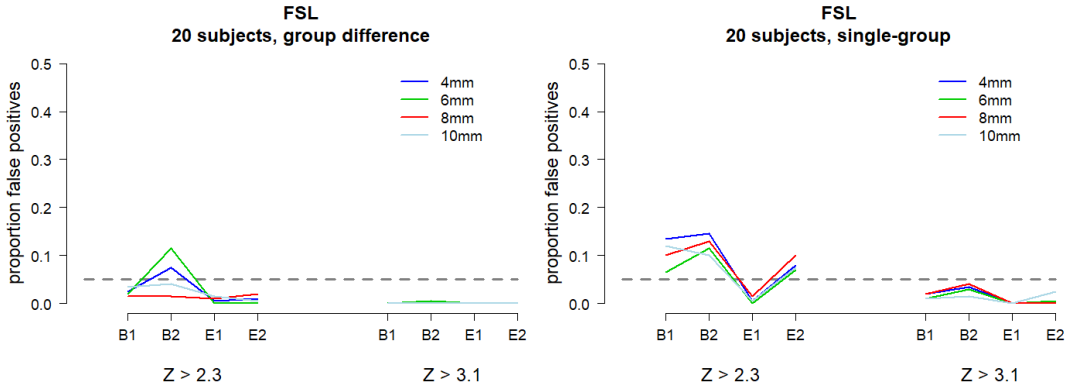
The ARI method found significant voxels in 6 scans in comparison to 992 scans with significant activity found by FSL. Table 3.2 describes the different settings of the scans that were found active by ARI.

Smoothing	**Design**	**Z-Threshold**	**Group size**	**Permutation**	**Contrast**	**PTD**
8mm	B1	2.3	40	19	A mean	0.00281
8mm	B1	3.1	40	19	A mean	0.00281
10mm	E2	2.3	20	31	A mean	0.00812
10mm	E2	3.1	20	31	A mean	0.00812
10mm	E2	2.3	40	31	A > B	0.02793
10mm	E2	3.1	40	31	A > B	0.02793

Table 3.2: Overview of the settings of the fMRI scans that were found significant after the ARI analysis including the smoothing levels, design, z-threshold, number of subjects within the group, random permutation number, contrast and the PTD.

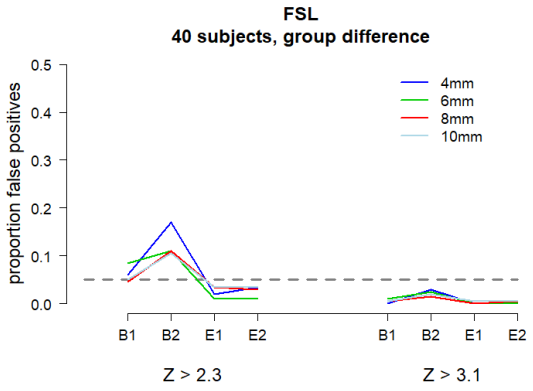
Table 3.2 shows that four of the significant scans were found by ARI during the 31st random permutation. The cluster forming thresholds were not used during the ARI analysis, but the results are mentioned twice in 3.2 to compare it with the FSL results. These four scans had a smoothness level of 10 voxels and the design E2. Two of them had an analysis level with 40 subjects and two groups for which the contrast A >group B was used. The other two scans had an analysis level with 20 subjects and one group with the mean of group A as contrast. The other two significant scans were during the 19th random permutation with a smoothness level of 8, design B2, 40 subjects, and one group. The significant differences were found in the contrast group A mean contrast.

The significant results of the FSL analysis are visualized in figure 3.1. The false positive proportion is visualized for all four different analysis levels and the two different z-thresholds.

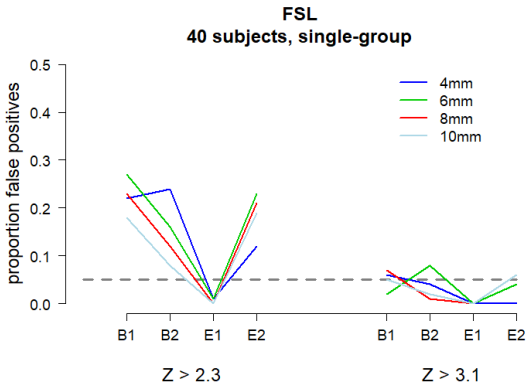


(a) 20 subjects, two groups

(b) 20 subjects, one group



(c) 40 subjects, two groups



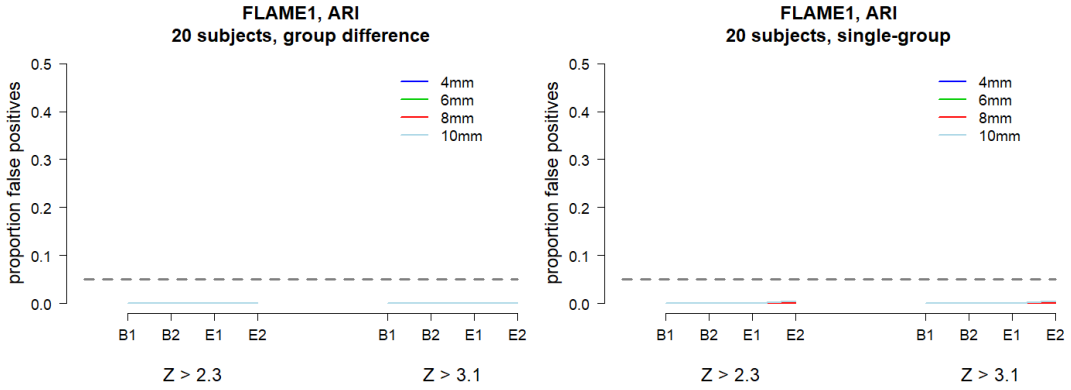
(d) 40 subjects, one group

Figure 3.1: Results of all different settings of the FSL analysis of null data. Each sub-image shows the results for different analysis levels. The y-axis describes the proportion of false positives. The x-axis represents two different variables. There are four designs on the x-axis for each of the z-thresholds. The different colors in the legend describe the size of the full width at half maximum (FWHM) used for smoothing.

The first image in figure 3.1a shows the results of the group difference analysis with 20 subjects in the two groups. Figure 3.1b shows the results of the single group analysis level with 20 subjects in the group. The third image in 3.1c represents the group difference between two groups of 40 subjects. The last image 3.1d is a visualization of the single group mean test with 40 subjects in the group. The z-threshold of 3.1 returns the lowest false positive rate across all four analysis levels. The proportion of false positives lies around zero for the design E1 for all different analysis levels and is the lowest of all four designs. The performance of the smoothing levels differs across other settings, but the smoothing levels of 4mm and 6mm perform on average worse compared to the 8mm and 10mm.

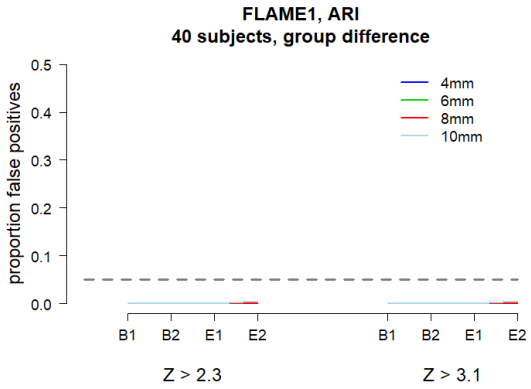
Figure 3.2 shows the false positive proportions of the ARI analysis with a sub-figure for each of the four analysis levels. There were only six significant scans during the ARI analysis, so the proportion of false positives is very low. Therefore, there are false positive rates of zero in most combinations of the preprocessing settings, except for the conditions in which activity was found. This is the reason that there is only some representation of the results in figures 3.2b, 3.2c and

3.2d.

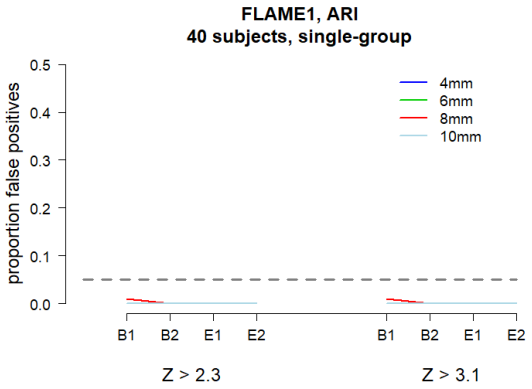


(a) 20 subjects, two groups

(b) 20 subjects, one group



(c) 40 subjects, two groups



(d) 40 subjects, one group

Figure 3.2: Results of all different settings of the ARI analysis of null data. Each sub-image shows the results for different analysis levels. The y-axis describes the proportion of false positives. The x-axis represents two different variables. There are four designs on the x-axis for each of the z-thresholds. The different colors in the legend describe the size of the full width at half maximum (FWHM) used for smoothing.

3.4 Discussion

The results show that a z-threshold of 2.3 finds much more false positives compared to a z-threshold of 3.1 during the FSL analysis. This indicates that research that performed their analysis with a low threshold might not be reliable in their findings. The activity that was found in fMRI scans when a low threshold was used could be false positives. However, the ARI method returns only 6 false positives compared to 992 false positives of FSL. This is in line with the hypothesis that the ARI method will perform better than FSL and will not find activity in many fMRI null data with no signal. However, the false positive rate of the ARI method is only 0.02% and is much lower than the 5% significance boundary that we expected. This could indicate that the ARI method does not have enough power to find activity in scans of subjects that were exposed to scans. Therefore, the next chapter will investigate the power of the ARI

method for scans that are exposed to tasks. However, Eklund et al. (2016) also found a low false positive rate for the null data that were pre-processed with FSL FLAME1. They found that the cause of this conservativeness was the zero between-subject variance. This indicates that the preprocessing had a big influence on the results. Therefore, a suggestion for further research is to redo the experiment with less conservative settings for null data in the preprocessing. This might lead to a false positive rate of around 5%. The results of the FSL analysis show that settings can influence the outcome of finding activity in scans. Because of the lack of activity found in the ARI method, conclusions about the influence of settings on the ARI results is not possible. The next section will investigate the power of the ARI method.

Chapter 4

Performance of the All-Resolutions Inference method in identifying activity on fMRI scans from the Neurovault database.

4.1 Introduction

The results of the previous section showed that there were only findings of activity in 6 of the 28800 scans when All-Resolutions Inference (ARI) was used. This led to uncertainty about the power of the ARI method. However, previous research did find the support that the ARI method has high power (Goeman et al., 2017; Rosenblatt et al., 2017). However, previous research did not compare the results of the ARI method to previous findings on the same data sets. Therefore, this chapter will answer the question of whether the ARI method will also find activity in scans exposed to stimuli when the accompanying articles of the data find activity. To answer this question, data from an online database called Neurovault is used. This database is easy to download and consists of a lot of available scans. The ARI method will be performed on this data set with the whole mask used as single cluster. The performance of the ARI method will be compared to the results found by the corresponding articles of the data. The hypothesis is that the ARI method will find activity in most scans when there was activity found in the article accompanying the data and sufficient thresholding was used. In other words, the expectation is that the ARI method has enough power to find activity in fMRI scans that are exposed to stimuli.

4.2 Materials and Methods

Data

As previously mentioned, the data that was used for the calculation of the power of the ARI method was gathered from the online database called Neurovault. An R-script was used to extract all collections available in this database through an API. However, the script was not consistent in its output. Each time the script was run, a different set of collections was downloaded and saved. Therefore, the extraction of the collections was done seven times until there were no new collection id's added. A collection can consist of multiple scans that are uploaded by the same user. The complete database existed of 380 collections after extracting the data seven times.

These collections contained very different types of data including MEG scans, resting state fMRI scans and task-based fMRI scans. The scans were represented by statistical z-maps, t-maps, and f-maps. This chapter is investigating the capability of the ARI method to find activity in scans during which tasks are performed, so only the task based scans need to be kept. Therefore, the collections were filtered on specific characteristics to delete the useless data for this research question. Therefore, a list was made with attributes on which the data needed to be filtered which are displayed in table 4.1.

Filter	**Description**	**Collections removed**
Statistical Maps	Only the scans that were T-maps or Z-maps were selected	170
Modality	Only fMRI-BOLD scans were selected	29
Activity	Resting state scans were removed	24
DOI	Collections that have a doi were selected	82
Subjects	2 collections with a disproportionate size (∼50% of total scans) were removed	2
Missing values	Remove missing values for discoveries	3

Table 4.1: Filter characteristics for the Neurovault data.

Table 4.1 shows that the filter steps include selecting T-maps and Z-maps of fMRI-BOLD scans that are linked to a doi article, removing resting state scans and removing two collections that contained half of the total number of scans. Only scans with a doi were selected because the doi accompanying the scans is important to compare the ARI results with the results described in the article accompanying the data. It is also important to look at the thresholding that was used in the article. We selected only the T-maps and Z-maps because these are the most often used statistical maps and these values can be used to calculate p-values from that are needed in the ARI analysis. There were two collections that had so many scans that they covered half of the available scans. The ARI method found activity in almost all of the scans from these two collections. This could give a power with a misleading high value as a result if these two collections were not removed.

After performing all the filter steps mentioned in table 4.1, the final data set consisted of 70 collections with a total of 820 images. The ARI method was run on this data set in R with the package ARIbrain (Rosenblatt et al., 2017). A collection often consists of multiple scans as mentioned before, so one doi covers multiple scans of the same collection. The collections in which no activity was found by the ARI method were investigated and the cluster forming thresholds were notated. In the articles of the collections were often less significant results discussed than the number of scans that were available in the collection. This indicates that the researchers themselves did not find any activity in some scans, but did upload those scans on Neurovault. If these scans are not removed and the ARI method does not find any activity in these scans as well, the power could turn out lower than it should be. These scans should not be taken into account while calculating the power. To control for this, 32 articles for which the ARI method did not find activity were randomly selected and the findings discussed in the article were observed. These 32 articles covered 401 images. The scans in which no significant activity was found in the article and the scans that were not discussed in the article were removed, these were 127 scans in total. The new power was calculated on the new final data set which consisted of 274 images. The results will be discussed in the next section.

4.3 Results

The ARI analysis on the scans that remained after filtering found activity in 72.1% of the scans. Among these scans, there were 44 collections in which the ARI method did not find activity in at least one image. The articles of these collections were further inspected to get a better image of the thresholds and analysis methods used in these articles. The articles often mentioned findings where the analysis was done with uncorrected thresholds or low cluster-forming thresholds. For example, Bohon (2017) used cluster-based thresholding in FSL with $Z > 1.7$. This is a very low threshold, which was chosen because it was a pilot study with a small sample size. This information was not included in the information given by Neurovault and could therefore not be filtered out. However, this does explain the lack of activity found by the ARI method in some scans. Table 4.2 shows the number of collections in which the ARI method did not find any discoveries in at least one image for each z-value threshold mentioned in the articles. Most collections have a cluster forming threshold lower than 3.1. Eklund et al. (2016) showed that there are more false positives for a cluster forming threshold of $z > 2.3$ compared to $z > 3.1$. The ARI method might not have found activity in these scans because there is actually no significant activity and the articles found false positives.

Threshold	**Number of collections**
$Z > 1.7$	2
$Z > 2.3$	19
$Z > 2.5$	1
$Z > 2.6$	9
$Z > 3.1$	13

Table 4.2: Number of collections that have different z-value thresholds in which the All-Resolutions Inference method did not find any discoveries in at least one scan within the collection.

After investigating the cluster forming thresholds, the significant findings in 32 random selected articles of the 44 articles were investigated. The scans for which no activity was discussed in the article were removed which improved the power of ARI and resulted in activity found in 78.1% of the scans.

4.4 Discussion

The final power of 0.78 is a good indication that the ARI method is indeed able to find activity in fMRI scans in which subjects were exposed to stimuli. This is in line with our expectations. The power will probably improve when the images that were analyzed in the article with a low threshold or uncorrected threshold are removed. Much of the Neurovault database consisted of images that were not useful for this research and decisions concerning the filtering of the data were made. These filter steps removed a big proportion of the images, but it was impossible to remove all faulty data. Some collections could have contained false information in their details and a lot of collections missed information that could have helped with the filtering. The real power might, therefore, be different from the power we found in the results. However, the database did include a lot of different designs, preprocessing settings, software programs, scanners etcetera. This is a wide range of possible influences on the ARI results that were all covered by the Neurovault database. Cleaning up the Neurovault database or finding another cleaner database is a good

next step for further research to get more precise results. This research also compared only 32 articles of the collections with the ARI results. Comparing the articles of all the collections with a doi with the ARI results will give a better estimation of the power. Another method that will get a more accurate power for the ARI method must be performed in further research.

Chapter 5

Building an app to support researchers in their use and experimentation with the All-Resolutions Inference Method.

5.1 Introduction

The previous chapters describe the performance of the ARI method on fMRI scans of subjects while they were exposed and unexposed to stimuli. The results show that the ARI method has a low false positive rate and has power to find activity in fMRI scans. However, no details have been given on how to perform the ARI method. Therefore, this chapter will discuss the application of the ARI method. The package ARIbrain was developed to apply the ARI method on fMRI scans from the command line in R (Rosenblatt et al., 2017). After applying the method, a list of information about the clusters will be returned. This includes information about the number of active voxels within the cluster, the number of inactive voxels within the cluster, the size of the cluster, the proportion of true discoveries (PTD), the dimensions and the test statistic. This sounds easy, but there are many different settings and different approaches to the application of the ARI method. A researcher can decide to use atlases as cluster maps or decide to drill down on specific clusters after running the ARI method once. Since the method is exploratory, the researcher should be able to try out different choices. This can all be done while maintaining proper FWER control. Most standard software programs do not allow a flexible approach to the analysis of fMRI scans. Besides that, there are no software programs that have the option to perform the ARI method yet. An app that will guide the user through the different approaches and settings can be very useful for researchers that have no experience with this method. Therefore, it is important to develop a flexible program that supports the ARI method and can guide researchers. Aside from the added value of guiding researchers through the analysis, the results will also be easier to interpret when the clusters are visualized on an image of the brain. This gives an indication of the location of the active clusters and also gives the names of the brain areas that fall inside a cluster. For example, researchers that have no prior indication of the location of activity caused by a specific task can form clusters based on the statistics values. They can start with a low threshold and zoom in on clusters that have a large number of active voxels. This will give an indication of the different locations of the activity. If a researcher does have a hypothesis on the brain area that is activated by a specific task, the clusters can be based on an atlas of the brain. After performing the ARI analysis, the researcher can then zoom in on the brain area of interest and look into the expectations. Both approaches can be combined with each other and can be performed on the same data. This is

all combined in an app that will be discussed in this chapter.

5.2 Materials and Methods

The app was build in R with the package shiny (Chang et al., 2018). The app is suitable for Z-maps and T-maps of task-based fMRI images. A shiny app is programmed into two parts, the server part, and the interface part. The server part includes all the functions and algorithms which are stored in some output variables. This output is displayed in the interface part. The interface part also takes in input from the user when a button is clicked or a selection is made in the app. This input is transferred to the server part which uses this information to adapt the output. Finalizing the server and the interface gives an interactive app.

How the app should look like

As mentioned in the introduction, the app is meant to guide the user in an exploratory analysis of fMRI scans. This means that one of the most important features of the app should be flexibility. Flexibility can be provided by allowing different settings and approaches which will be built into the app. Another feature of the app will be the visualization of the results. FMRI scans are acquired in axial, coronal and sagittal planes. These planes together represent a 3D image of the brain which is important for the visualization and localization of brain activity. To make a good visualization of the image, all three planes will be plotted in the app. The MNI152 standard-space T1-weighted average structural template image from FSL with a slice thickness of 2mm will be used as a standard background on which the atlases and results of the ARI method will be displayed. The atlases that will be used in the app are also gathered from FSL. Therefore, each user should have downloaded the FSL folder in which those images are stored. The user will have to upload the file path towards the FSL folder on their computer in the app to get access to those images. The user will also be able to upload multiple files to perform the ARI analysis. The most important thing is the statistics map of the fMRI image which can only consist of z-values or t-values. These are the only maps allowed because p-values can be calculated from these maps. Another option will be to upload the p-values map, but the app should be able to calculate the p-values from the z-values or t-values themselves. To calculate the p-values and to run the ARI analysis, the user will have to define the type of statistic values (t-values or z-values) and if necessary the degrees of freedom. The possibility to upload a mask accompanying the statistic map will also exist. However, this is not required, since the mask can also be calculated from the non-zero values in the statistics map. This includes all the necessary information to run the ARI analysis. However, the user will be provided with the possibility of adding an atlas of the brain which will be visualized on the background image. This atlas can be used as a cluster map during the ARI analysis, but the user can also choose to base the cluster map on the statistics map. After uploading and making these decisions, the user will be able to choose a z-value threshold for the ARI method. Then the analysis is run and the resulting clusters will be displayed on the three planes on the dashboard page. A progress bar will keep track of the progress of the analysis. If the user clicks on a cluster, information about that cluster will appear next to the images in the right bottom. If an atlas was selected, the brain area in which the selected voxel lies will also appear. Information about all the clusters in a data frame after the ARI analysis was run will appear in a new tab on the dashboard page. The user will be able to zoom in, also called drill down, on specific clusters, spheres and brain areas after the ARI analysis. The user will also be able to form new clusters by growing them from local optima. A progress bar will keep track of the progress of growing the clusters from local optima.

Zooming in on a cluster is defined as selecting resulting clusters after the ARI analysis that can be adjusted by changing the threshold as deemed necessary. The cluster will become smaller if the threshold is set to a higher value and bigger if the threshold value is set to a lower value. This can be useful when a cluster covers a lot of different brain areas and has a PTD of 0.5. The cluster will be divided into multiple clusters when the threshold is set to a higher value. This can result in smaller clusters with a higher PTD. Another possibility to zoom in is looking at a sphere around a voxel. A sphere is a three-dimensional circle around the selected voxel with a radius of a specific number of voxels. The app will allow the user to select the size of the radius. After defining the radius size, the threshold for the sphere will be set to select only the voxels within the sphere that are above that threshold value. The third option is to zoom in on brain areas that are based on the selected brain atlas. The app will provide the possibility to the user to choose a brain area for which the threshold needs to be changed. After selecting this brain area and altering the threshold, the ARI method is run again for only the voxels within that brain area. The last option that needs to be included in the app is the possibility to grow clusters from z-value optima. The statistics file that was uploaded in the app has a statistical value for each of the voxels. The maximum z-value represents a voxel with the highest probability of being active. The user should be able to select the maximum size of the clusters that are grown, the number of clusters to grow and the threshold for these clusters. Another option is to select whether the user wants a cluster with the highest PTD found or a cluster that is bigger and does not go below a selected PTD threshold. The app will allow these options and an algorithm will be developed. The algorithm will begin with growing one cluster from the biggest z-value and stop when the surrounding z-values are all below the z-value threshold, when the cluster size gets above the maximum size or the PTD decreases. The cluster that returned the highest PTD is the eventual cluster that is saved. The voxels that are not included in the previous clusters are used for growing the next cluster. The maximum z-value among those voxels is the next optima. The original cluster map after running the ARI method will be removed and replaced with the new cluster map made by growing clusters from local optima. The results after each zoom in action will be saved in a new tab on the dashboard page. There will be several download buttons to download the images, the cluster map, the ARI results, and the zoom-in results.

5.3 Results

After building the interactive shiny app, the app was tested. The steps to follow within the app and the results will be discussed in this section. We will try out the application on a task-based fMRI scan. Image number 19155 from the neurovault database is used to describe the use of the application (https://neurovault.org/images/19155/). This image was chosen because the ARI method found a lot of active voxels in this image in the previous chapter. Therefore, this is a good example to use for demonstrating the application.

Step by step explanation on how to use the application

1. Open the app and maximize the screen to full size. You will now see the dashboard page.

2. There are several download buttons in the left sidebar of the app. These buttons can be used to download the images, the results of the ARI analysis, the zoom in analysis results and the cluster map. The plots of the brain images can be downloaded after loading in the files in step 1. The ARI results and cluster map can be downloaded after performing the ARI analysis in step 2. To download the zoom-in results, first, zoom in on voxels in step

3. The zoom-in results of the previous zooming in steps will be removed when the growing clusters from the z-value optima option are performed.

3. Click in the low right corner on: "Step 1: Click here to load the fsl directory, statistics and mask files" or click in the top left corner on 'Load in Files' to perform step 1.

   a) Click under the title 'FSL directory' on "Select your fsl directory" and browse to the local folder on your computer where fsl is downloaded. Select your fsl folder and click on 'Select' as shown in figure 5.1. Make sure you have downloaded fsl from the official website https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/. **You need to do this in order to try out the example.**

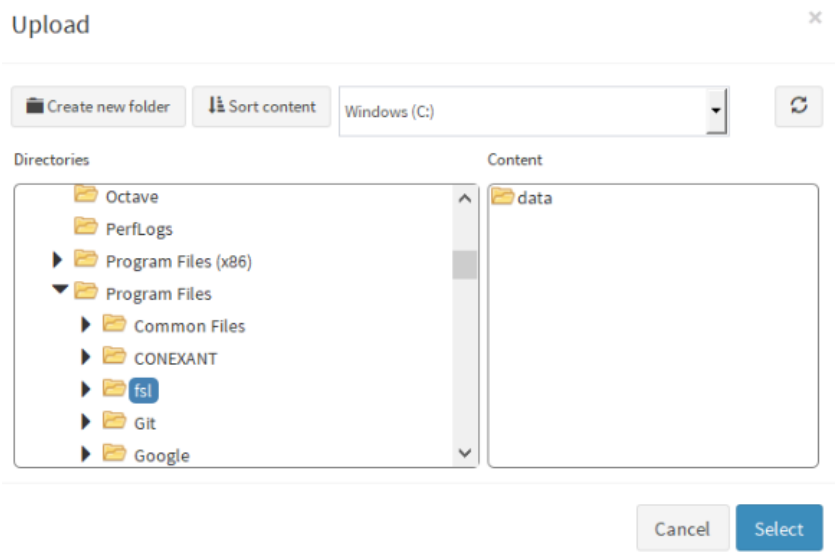


Figure 5.1: Loading the fsl folder to subtract the atlases.

   b) Under the title 'Statistics file' browse to the file of your z-map or t-map of your fMRI image that you want to analyze and open this file. **For the example, upload the image 19155 of Neurovault under the title 'Statistics file'.**

   c) Select whether you uploaded a map with z-values or t-values. If you uploaded t-values, insert the degrees of freedom. **The Neurovault image 19155 of the example is a map consisting of z-values.**

   d) If you have a p-values map of the statistics map, you can upload the file under the 'p-values of statistics (Not Required)' file. This is not required because the app is also able to calculate the p-values from the statistics map. **For the example, you do not need to upload p-values.**

   e) Select under 'Upload mask or use the statistics map as mask' whether you want to upload a mask you received after preprocessing the scan or use the non-zero values in the statistics map as mask. **Select the option 'Use map as mask' for the example.**

   f) You can choose an atlas under the title 'Choose an atlas' to see which brain areas have activity after running the All-Resolutions Inference (ARI) method. There are

multiple atlases that you can choose from: No atlas, Talairach, Harvard Oxford - Cortical, Harvard Oxford - Subcortical, Cerebellum, John Hopkins University (JHU), Juelich, Mars Parietal Parcellation, Mars Temporoparietal Junction Area (Mars TPJ) Parcellation, Montreal Neurological Institute and Hospital (MNI), Neubert Ventral Frontal Parcellation, Sallet Dorsal Frontal Parcellation, Striatum and the Thalamus. **Select the 'Talairach' atlas for the example.**

g) Make sure the settings on your screen are the same as figure 5.2

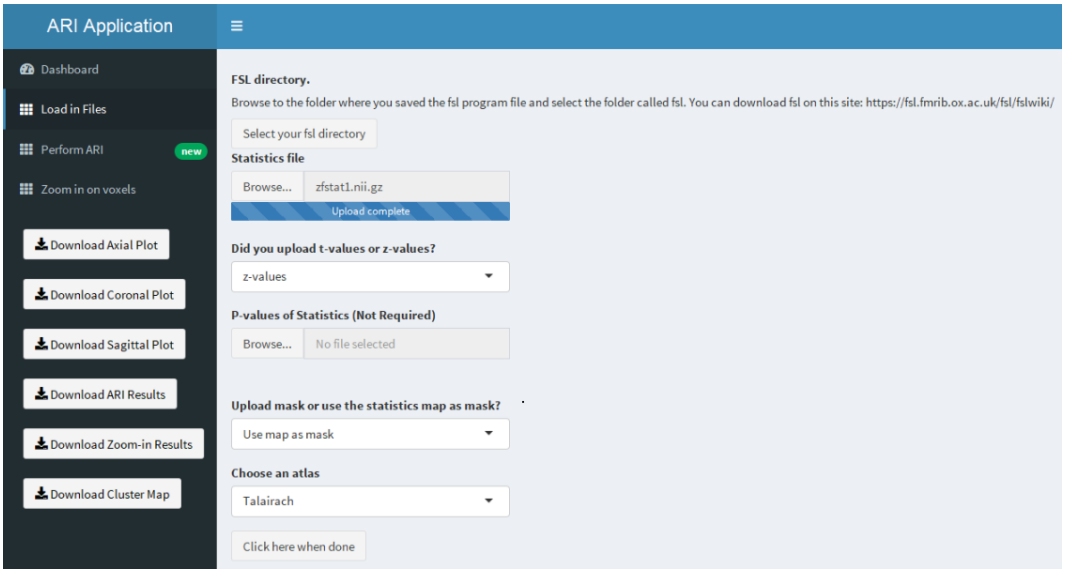


Figure 5.2: Loading in files.

h) After following the above steps, click on "Click here when done". The app will return to the dashboard page and will show the brain images.

i) Figure 5.3 shows how the dashboard page should look like when the files are correctly loaded. You can now download the plots of the brain images as displayed on the dashboard.

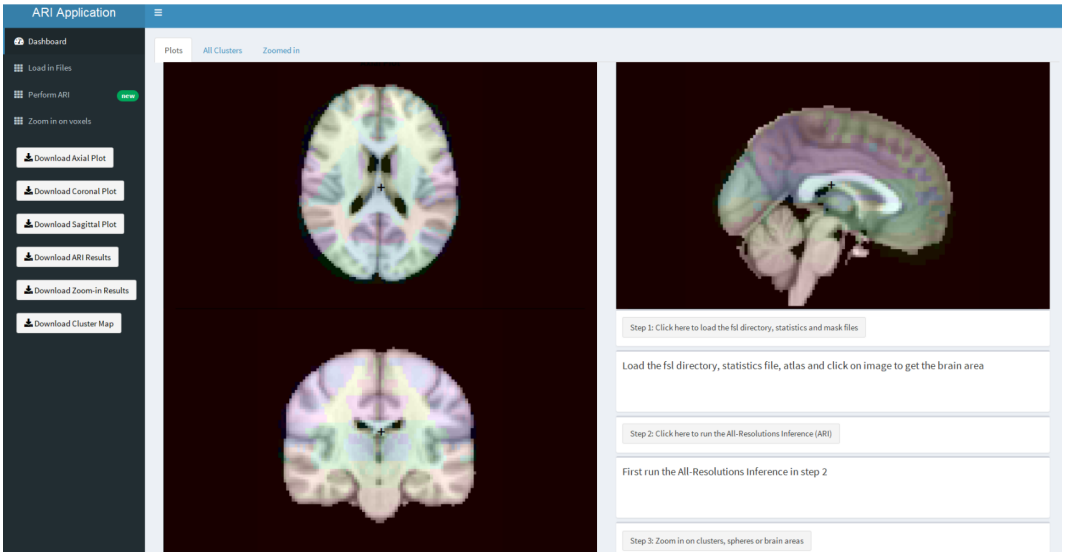


Figure 5.3: Dashboard after loading in files and selecting the Talairach atlas.

4. Click in the bottom right corner of the dashboard on "Step 2: Click here to run the All-Resolutions Inference (ARI)" or in the top left corner on: "Perform ARI" to run the analysis and perform step 2.

   a) The stat-threshold input bar represents values between the minimum z-value and the maximum z-value of the statistics map. The middle value is automatically selected. Choose the cluster threshold you want to use. **For the example, we use a threshold of 3.1.**

   b) Select under the heading "Select atlas to use as cluster map" the atlas you want to use as cluster map or use the statistics map as cluster map. An atlas will use each brain area as a separate cluster. **For the example, choose "Based on statistics map".**

   c) Figure 5.4 shows how the page should look like after following above steps.

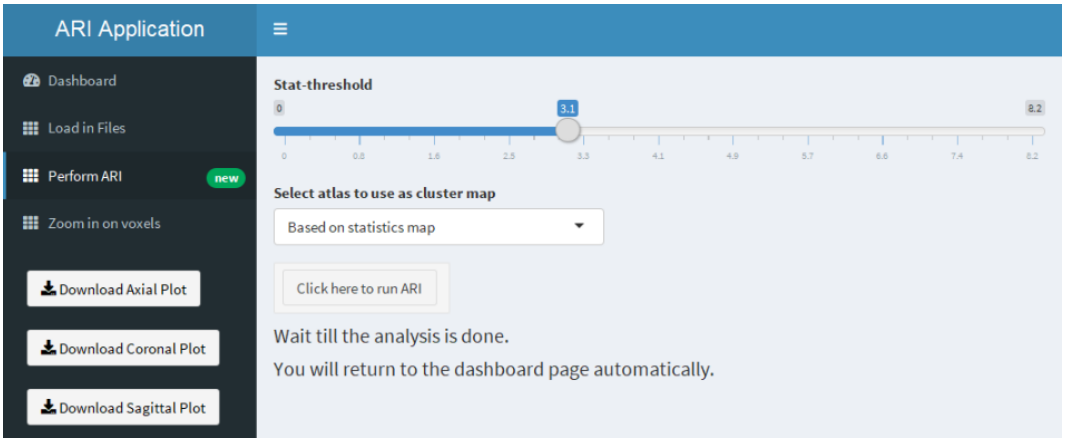


Figure 5.4: Settings for running the All-Resolutions Inference.

d) **Click on "Click here to run ARI" to run the analysis.** A progress bar will appear in the bottom right of the screen called "Running Analysis". The app will return automatically to the dashboard page after the analysis is done.

a) The clusters are displayed within the images on the dashboard page. Each cluster has a different color. If you click on a cluster, information is shown in the right bottom of the screen next to the images. **If you click on the red cluster in one of the images while doing the example, you should see around the same as in figure 5.5.** You can now download the plots of the images, the ARI results and the cluster map.

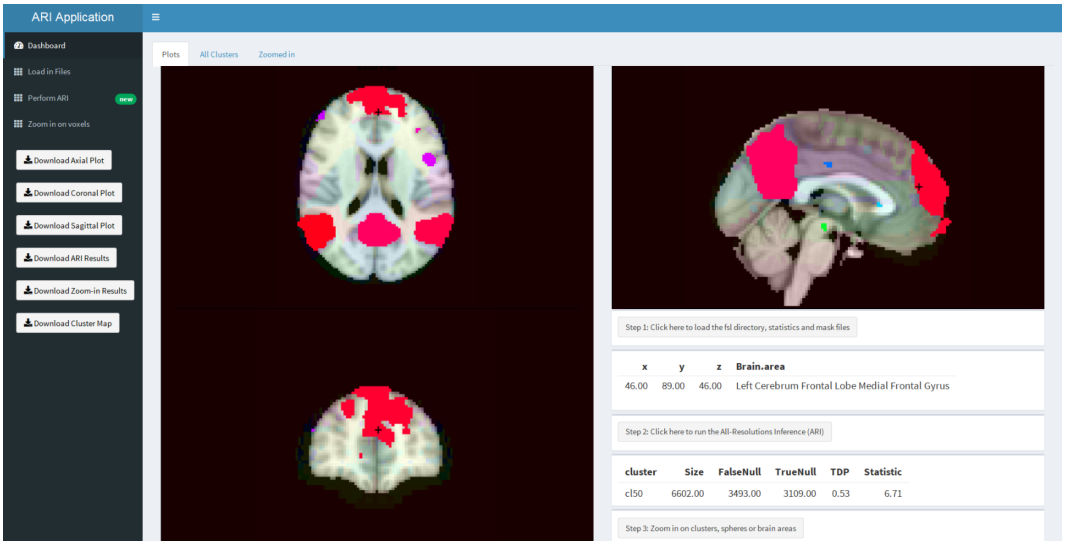
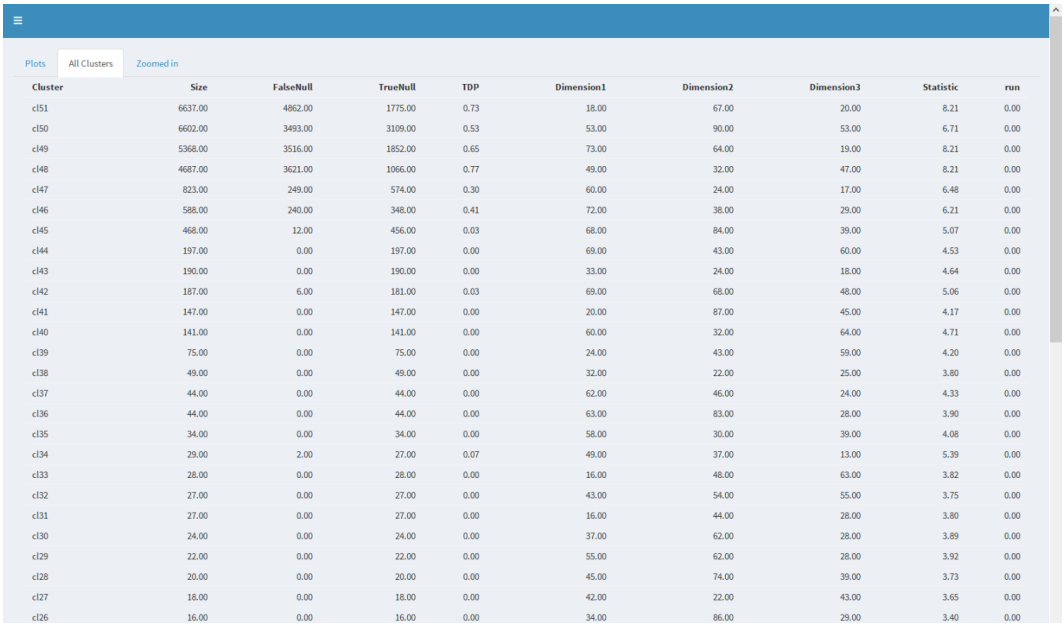


Figure 5.5: Dashboard after running the ARI method.

Information on the clusters include:

i. The name of the cluster
ii. The size of the cluster (number of voxels)
iii. The number of active voxels within that cluster (FalseNull)
iv. The number of inactive voxels within the cluster (TrueNull)
v. The proportion of true discoveries (PTD) which is the number of active voxels divided by the size of the cluster.
vi. The value of the maximum statistic value within that cluster.
vii. The brain area of the selected voxel within the image.

b) Click on the tab "All Clusters" within the dashboard page to get a list with information on all the clusters. **The page should look like figure 5.6 in the example.**



Plots | All Clusters | Zoomed in

Cluster	Size	FalseNull	TrueNull	TDP	Dimension1	Dimension2	Dimension3	Statistic	run
cl51	6637.00	4862.00	1775.00	0.73	18.00	67.00	20.00	8.21	0.00
cl50	6602.00	3493.00	3109.00	0.53	53.00	90.00	53.00	6.71	0.00
cl49	5368.00	3516.00	1852.00	0.65	73.00	64.00	19.00	8.21	0.00
cl48	4687.00	3621.00	1066.00	0.77	49.00	32.00	47.00	8.21	0.00
cl47	823.00	249.00	574.00	0.30	60.00	24.00	17.00	6.48	0.00
cl46	588.00	240.00	348.00	0.41	72.00	38.00	29.00	6.21	0.00
cl45	468.00	12.00	456.00	0.03	68.00	84.00	39.00	5.07	0.00
cl44	197.00	0.00	197.00	0.00	69.00	43.00	60.00	4.53	0.00
cl43	190.00	0.00	190.00	0.00	33.00	24.00	18.00	4.64	0.00
cl42	187.00	6.00	181.00	0.03	69.00	68.00	48.00	5.06	0.00
cl41	147.00	0.00	147.00	0.00	20.00	87.00	45.00	4.17	0.00
cl40	141.00	0.00	141.00	0.00	60.00	32.00	64.00	4.71	0.00
cl39	75.00	0.00	75.00	0.00	24.00	43.00	59.00	4.20	0.00
cl38	49.00	0.00	49.00	0.00	32.00	22.00	25.00	3.80	0.00
cl37	44.00	0.00	44.00	0.00	62.00	46.00	24.00	4.33	0.00
cl36	44.00	0.00	44.00	0.00	63.00	83.00	28.00	3.90	0.00
cl35	34.00	0.00	34.00	0.00	58.00	30.00	39.00	4.08	0.00
cl34	29.00	2.00	27.00	0.07	49.00	37.00	13.00	5.39	0.00
cl33	28.00	0.00	28.00	0.00	16.00	48.00	63.00	3.82	0.00
cl32	27.00	0.00	27.00	0.00	43.00	54.00	55.00	3.75	0.00
cl31	27.00	0.00	27.00	0.00	16.00	44.00	28.00	3.80	0.00
cl30	24.00	0.00	24.00	0.00	37.00	62.00	28.00	3.89	0.00
cl29	22.00	0.00	22.00	0.00	55.00	62.00	28.00	3.92	0.00
cl28	20.00	0.00	20.00	0.00	45.00	74.00	39.00	3.73	0.00
cl27	18.00	0.00	18.00	0.00	42.00	22.00	43.00	3.65	0.00
cl26	16.00	0.00	16.00	0.00	34.00	86.00	29.00	3.40	0.00

Figure 5.6: Information on the clusters.

c) Go back to the tab "Plots" on the dashboard page.

5. Click in the bottom right corner of the dashboard on "Step 3: Zoom in on clusters, spheres or brain areas" or in the top left corner on "Zoom in on voxels" to perform step 3 and change the threshold of a particular subset of voxels and run the ARI analysis again on only that subset.

   a) Choose between the following zoom in possibilities:

      i. Cluster: Select a cluster from the clusters formed in the initial ARI analysis and change the threshold. The cluster size will probably become smaller if the threshold is set to a higher value and bigger if the threshold value is set to a lower value.

      ii. Sphere. A sphere is a 3-dimensional circle of voxels around a selected voxel. Go back to the dashboard page and select the voxel in the image that you want to use as the middle point of your sphere. Then go back to the tab "Zoom in on voxels" and choose the sphere size. The sphere size represents the number of voxels in the radius of the sphere. Then set the threshold for the sphere to select only the voxels within the sphere that are above that threshold value.

      iii. Brain area. The brain areas are based on the atlas that you selected at the "Load in Files" tab. You have to run the initial ARI analysis again with a different atlas to select different brain areas. Choose the brain area for which you want to change the threshold. Next step is to change the threshold and run the ARI method again for only that brain area.

      iv. Grow clusters from z-value optima. The statistics file that was uploaded during step 1 has many z-values. The maximum z-value represents a voxel which is most likely to be active. The user can select the maximum size of the clusters, the

number of clusters to grow and the threshold for these clusters. The algorithm will begin with growing one cluster from the biggest z-value and stop when the surrounding z-values are all below the threshold when the cluster size gets above the maximum or the PTD decreases. You can choose whether you want the cluster that returned the highest PTD is the eventual cluster that is saved or a cluster that did not get below a specific PTD threshold. The voxels that are not included in the previous clusters are used for growing the next cluster. The maximum z-value among those voxels is the next optima. The original cluster map after running the ARI method is removed and replaced with the new cluster map made by growing clusters from local optima. The previous zoom-in results will be removed after this setting is applied.

b) After doing any of the zoom-in methods you can now download the plots of the images, the ARI results, the cluster map and the zoom-in results.

c) **For the example, we are going to zoom in with all the possible settings:**

i. Select 'Cluster' under the 'Select zoom in method' box. Choose cluster 51 with a threshold of 5.7. Make sure the settings are the same as in figure 5.7 and click on 'Select Voxel & Method, Change Threshold then Click Here'.

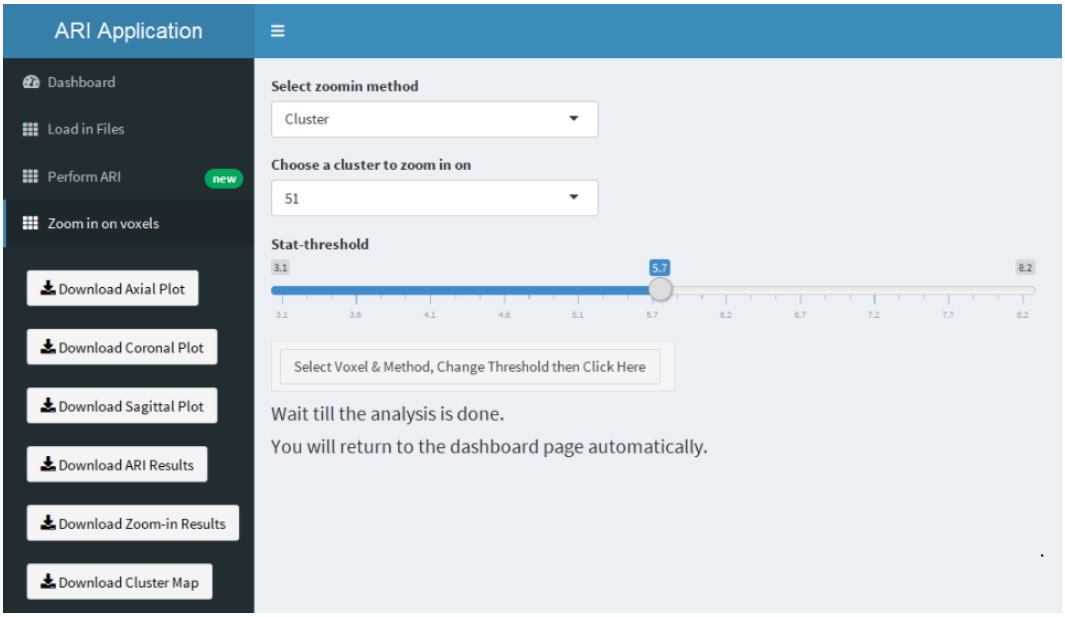


Figure 5.7: Settings for zooming in on the clusters.

ii. Go to the dashboard and choose a voxel within a cluster in the middle of the images and go back to the 'Zoom in on voxels' page. Set the sphere size to 15 and the threshold to 5 as is shown in figure 5.8 and click on 'Select Voxel & Method, Change Threshold then Click Here'.

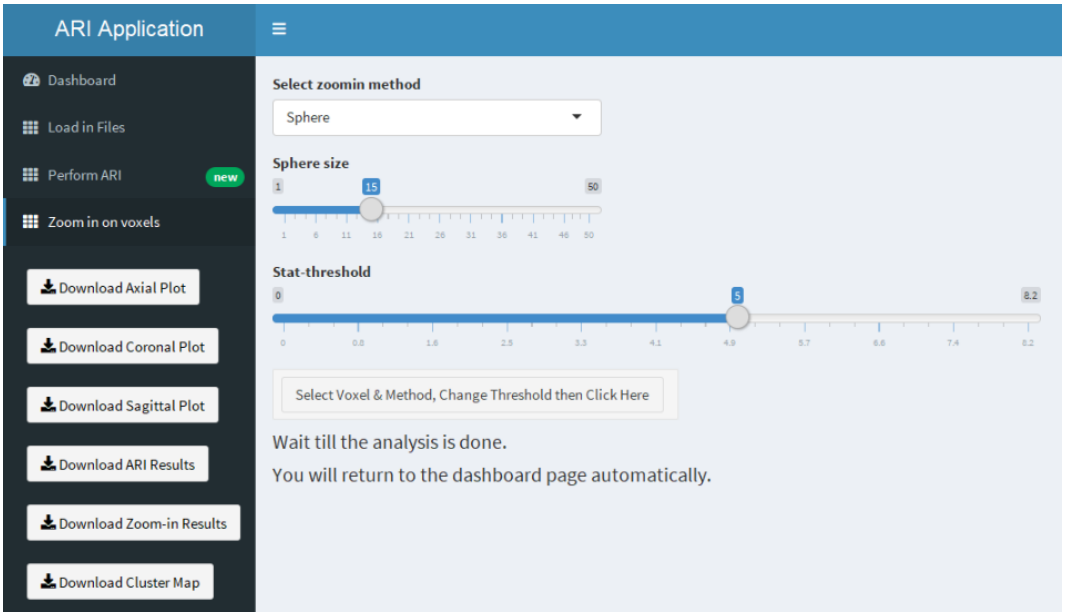


Figure 5.8: Settings for zooming in on a sphere around a selected voxel in the dashboard.

iii. Go back to the page of 'Step 3: Zoom in on clusters, spheres or brain areas' and select the option to zoom in on brain areas. Select from the Talairach atlas the brain area Left Cerebellum Posterior Lobe Inferior Semi-Lunar Lobule Gray Matter with a threshold of 4.1 and click on 'Select Voxel & Method, Change Threshold then Click Here'. The settings should be the same as displayed in figure 5.9. Then click on 'Select voxel & method, Change threshold then Click Here.'

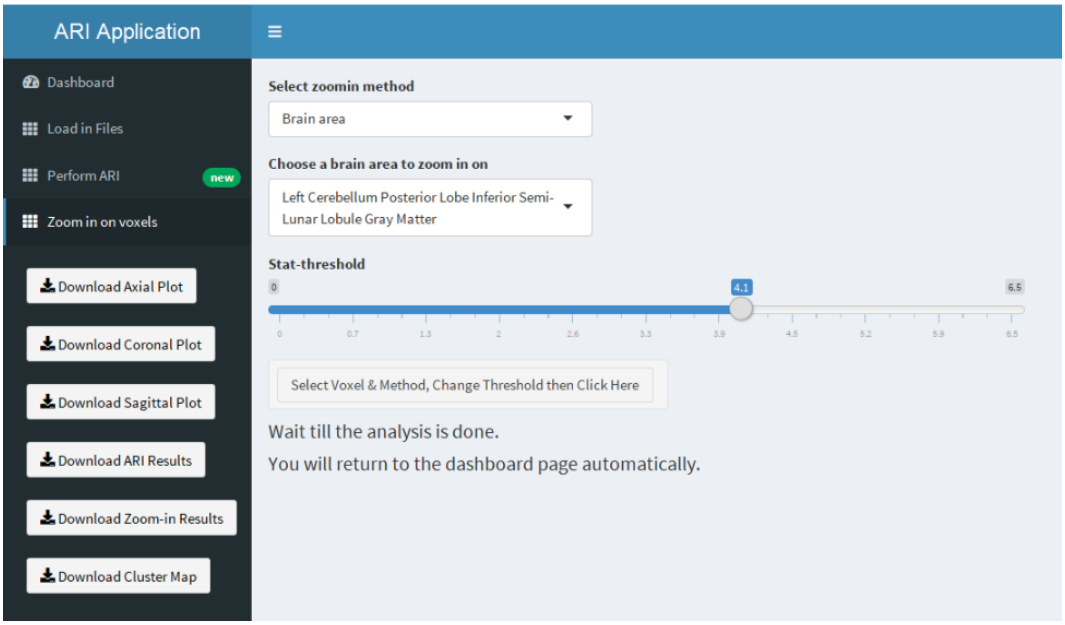
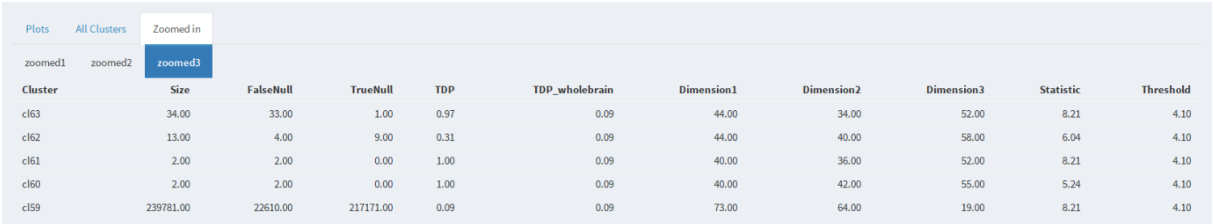


Figure 5.9: Settings for zooming in on a sphere around a selected voxel in the dashboard.

iv. Go to the tab 'Zoom in' from the dashboard to look at the ARI results of the different zoom in methods. This page should look like figure 5.10.



Cluster	Size	FalseNull	TrueNull	TDP	TDP_wholebrain	Dimension1	Dimension2	Dimension3	Statistic	Threshold
cl63	34.00	33.00	1.00	0.97	0.09	44.00	34.00	52.00	8.21	4.10
cl62	13.00	4.00	9.00	0.31	0.09	44.00	40.00	58.00	6.04	4.10
cl61	2.00	2.00	0.00	1.00	0.09	40.00	36.00	52.00	8.21	4.10
cl60	2.00	2.00	0.00	1.00	0.09	40.00	42.00	55.00	5.24	4.10
cl59	239781.00	22610.00	217171.00	0.09	0.09	73.00	64.00	19.00	8.21	4.10

Figure 5.10: Different tabs for the results of the zoom in methods.

v. The last option of zooming in is growing clusters from z-value optima. Select the zoom in method 'Grow clusters from z-value optima' and choose to grow 1 cluster. Set the threshold to 5.7 as shown in figure 5.11. Then click on grow optima. A progress bar of the analysis appears on the bottom right corner of the screen. Be aware that this analysis may take some time.

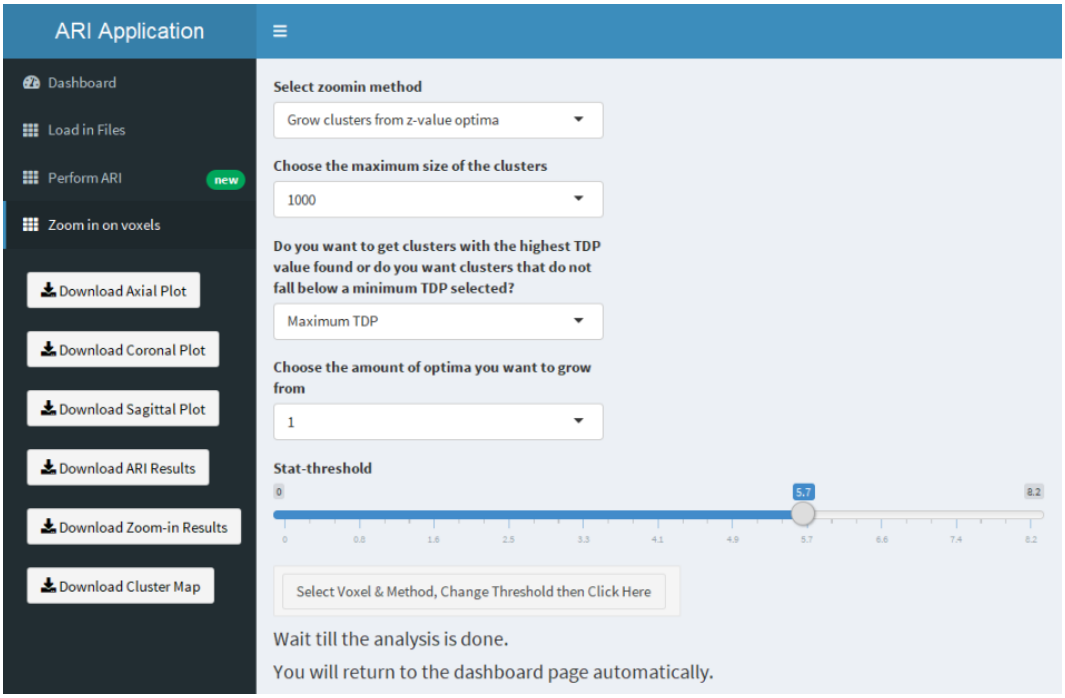


Figure 5.11: Different tabs for the results of the zoom in methods.

vi. After the growing of the optima is done, you will return to the dashboard page automatically which should look like figure 5.12.

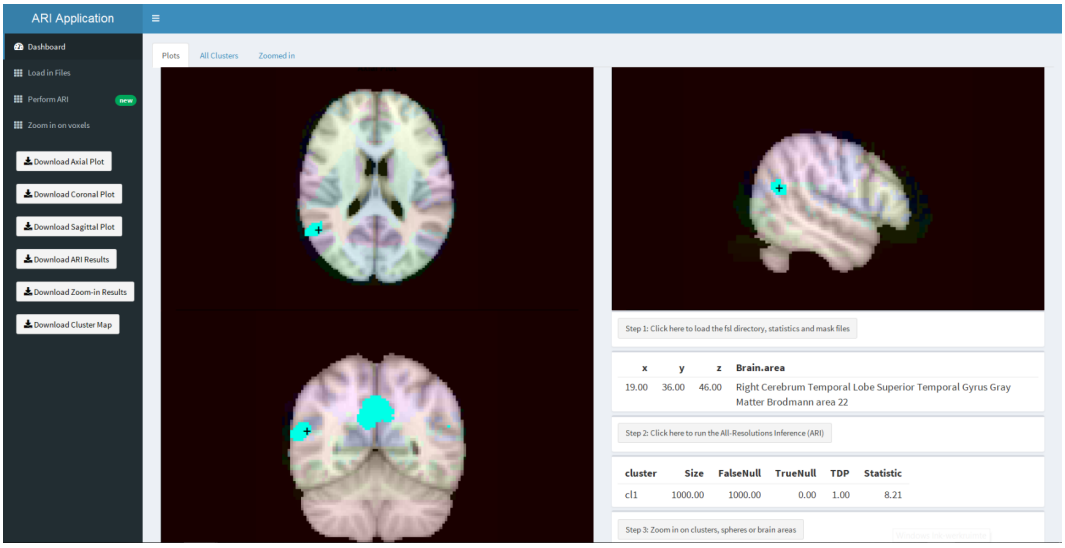


Figure 5.12: Results of growing a cluster from one z-value optima with a cluster threshold of 5.7.

d) After zooming in with the methods discussed above, the resulting clusters can be found in the tab "Zoomed in" on the dashboard page. Within the zoomed in tab are

the clusters displayed that are made by zooming in. A new tab is made for each zoom in method. You can zoom in multiple times.

5.4 Discussion

The app is a good support tool for researchers in performing the ARI analysis. It allows for an exploratory approach and provides a lot of flexibility which are the main advantages of the ARI method. This flexibility is shown during the example in the previous section. The example shows that cluster 51 of the ARI results has a proportion of true discoveries of 0.75. This means that 75% of the voxels in that cluster are active. There are five other clusters that have a TDP higher than 0 shown in figure 5.6. In the example, only cluster 51 was zoomed in on. For further investigation of active brain areas, the user can zoom in on the other clusters as well. The PTD increased after changing the threshold of cluster 51 and this is also expected for the other clusters. When the option to grow clusters from local optima was selected, a cluster with a PTD of one as shown in figure 5.12 was made. This means that all voxels inside that cluster are active. The example showed that there are a lot of different settings possible inside the app. However, there are still many improvements that can be made. One problem of the analysis is that the ARI method needs to calculate all the distances between the statistic values to generate clusters. The function 'dist' from the package 'stats' returns a vector that allocates a lot of memory (R core team, 2017). The most fMRI images do not give any problems while running the cluster forming algorithm, but some statistics maps have a lot of values above the selected threshold which will lead to the calculation of many distances. Therefore, computers need sufficient free memory to be able to analyze these big maps. Another improvement for the app would be to make the algorithms in the app faster. For example, the algorithm that uses local optima to grow clusters or the algorithm that gathers all the brain areas of a cluster. The app does keep track of the progress of the analysis which is a good indication of the time the analysis will take. In terms of reproducing an ARI analysis that is performed by the app, an overview of all the steps it took should be given. This overview can include information about the different thresholds or cluster maps selected and the different zoom in methods on which voxel sets were used. A log of the code that was used by the app would also be useful for the reproducibility. Further research can add these improvements to the app.

Chapter 6

Conclusion

Previous research already found that the ARI method is a big improvement on other analysis methods because of the exploratory nature and the flexibility of the ARI approach (Goeman et al., 2017; Rosenblatt et al., 2017). This study answered outstanding questions about the ARI method and found support for the ARI method in terms of high power and low false positive rates.

Chapter 3 showed that the All-Resolutions Inference (ARI) method has a low false positive rate and performs better than the software program FSL which was in line with the expectations. This was investigated by running ARI and FSL on the same fMRI null data where no activity should be found. However, the false positive rate was lower than the expected 5% which could indicate low power. Previous research already found that the power of ARI was high, but the ARI results were not compared with results found on the same data by published articles (Goeman et al., 2017; Rosenblatt et al., 2017).

The power of the ARI method was compared to the results of existing articles in chapter 4. The images used for the power calculation were downloaded from the online database Neurovault. A lot of collections within that database were useless for this purpose and therefore filtered out. Eventually, images of 32 articles were used to calculate the power which resulted in a power of 0.781. This is an acceptable power which indicates that the ARI method can find activity in a task-based fMRI scan. To investigate the power even more, the articles of the scans that remained after filtering out the useless images were investigated. Some articles of the scans in which no activity was found by the ARI analysis used uncorrected thresholds or very low cluster thresholds. The power might be higher when these images are removed or a cleaner database is used. Therefore, further research should be done to make a better estimation of the power.

The first two chapters give many indications that the ARI method is well suited for fMRI analysis. The method is however, not yet implemented in any existing software program since the method is newly developed and most software programs do not allow an exploratory approach. The exploratory nature is a big advantage of the ARI method in comparison to other analysis methods. Flexibility is therefore very important in the application of the ARI method. An app that was designed to be flexible and to allow many different approaches and settings to provide an exploratory analysis was therefore developed in chapter 5. This app is now ready to support researchers that are willing to use the ARI method. However, there are still many improvements to be made to the app. One improvement is to support the reproducibility of the analysis. The user should be able to look at the different steps that were taken and the output of the ARI method. The app is the first prototype and researchers will probably find bugs when they use

the app. A good next step is therefore to ask researchers to start using the app and give feedback on the problems they run into.

Besides the proof that the ARI method is good in finding activity in fMRI scans discussed before, there are some limitations to the method. Rosenblatt et al. (2017) found that the power of the ARI method decreases when the size of the cluster decreases. The detection of a single voxel has therefore the lowest power. This means that the ARI method is better in detecting activity in big areas than in small active areas within the brain. This is why the ARI method is only able to say how many voxels in a region are active, but not which voxels are active. For example, if the activity in a big cluster is scattered across the whole cluster, drilling down might not be possible.

The advantages of the ARI method exceed the limitations of the method. It is overall a great improvement on the existing analysis methods for fMRI research. The main advantage of the ARI method is that it allows an exploratory approach. Other analysis methods do not provide this possibility. The ARI method is supported by the power that was found in this study and the low false positive rate. The app that was made available for researchers to perform the exploratory ARI method is a great support in making the method practically accessible.

Bibliography

[1] Bandettini, P.A., Jesmanowicz, A., Wong, E.C. and Hyde, J.S. (1993). Processing strategies for time-course data sets in functional MRI of the human brain. *Magnetic Resonance in Medicine, 30*(2), 161-173. doi:10.1002/mrm.1910300204

[2] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological), 57*(1), 289-300.

[3] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics, 29*(4), 1165-1188.

[4] Biswal, B.B., Mennes, M., Zuo, X.N., Gohel, S., Kelly, C., Smith, S.M., Beckmann, C.F., Adelstein, J.S., Buckner, R.L., Colcombe, S., Dogonowski, A.M., Ernst, M., Fair, D., Hampson, M., Hoptman, M.J., Hyde, J.S., Kiviniemi, V.J., Kötter, R., Li, S.J., Lin, C.P., Lowe, M.J., Mackay, C., Madden, D.J., Madsen, K.H., Margulies, D.S., Mayberg, H.S., McMahon, K., Monk, C.S., Mostofsky, S.H., Nagel, B.J., Pekar, J.J., Peltier, S.J., Petersen, S.E., Riedl, V., Rombouts, S.A., Rypma, B., Schlaggar, B.L., Schmidt, S., Seidler, R.D., Siegle, G.J., Sorg, C., Teng, G.J., Veijola, J., Villringer, A., Walter, M., Wang, L., Weng, X.C., Whitfield-Gabrieli, S., Williamson, P., Windischberger, C., Zang, Y.F., Zhang, H.Y., Castellanos, F.X., Milham, M.P. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences of the United States of America, 107*(10), 4734-4739.

[5] Bohon, C. (2017). Brain response to taste in overweight children: A pilot feasibility study. *Plos One*. doi:10.1371/journal.pone.0172604

[6] Chang, W., Cheng, J., Allaire, J.J., Xie, Y. and McPherson, J. (2018). Shiny: Web Application Framework for R. *R package version 1.2.0.* https://CRAN.R-project.org/package=shiny

[7] Chi, Z. (2007). On the performance of FDR control: constraints and a partial solution. *The Annals of Statistics 35*(4), 1409–1431.

[8] Choong-Wan, W., Krishnan, A. and Wager, T.D. (2014). Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *NeuroImage, 91*, 412-419. doi:10.1016/j.neuroimage.2013.12.058

[9] Eklund, A., Nichols, T.E. and Knutsson, H. (2016). Cluster failure: Why fMRI infereneces for spatial extent have inflated false-positive rates. *PNAS, 113*(28), 7900-7905. doi:10.1073/pnas.1602413113

[10] Finner, H.,M. Roters, and K. Strassburger (2014). On the Simes test under dependence. *Statistical Papers*, 1–15.

[11] Goeman, J.J., Meijer, R.J., Krebs, T.J.P. and Solari, A. (2017). Simultaneous Control of All False Discovery Proportions in Large-Scale Multiple Hypothesis Testing. *arXiv preprint arXiv:1611.06739*

[12] Goeman, J.J. and Solari, A. (2011). Multiple Testing for Exploratory Research.*Statistical Science, 26*(4), 584-597. doi: 10.1214/11-STS356

[13] Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika 75*(2), 383–386.

[14] Marcus, R., Peritz, E. and Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika, 63*(3), 655-660. doi:10.2307/2335748

[15] Meijer, R.J., Krebs, T.J.P., Solari, A. and Goeman, J.J. (2017). Simultaneous Control of All False Discovery Proportions by an Extension of Hommel's Method. *arXiv preprint arXiv:1611.06739v1*

[16] Nichols, T. and Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: A comparative review. *Statistical Methods in Medical Research, 12*(5), 419-446.

[17] R Core Team (2017). R: A language and environment for statistical computing. *R Foundation for Statistical computing.* URL https://www.R-project.org/

[18] Rødland, E. A. (2006). Simes' procedure is 'valid on average'. *Biometrika 93*(3), 742–746.

[19] Rosenblatt, J.D., Finos, L., Weeda, W.D., Solari, A. and Goeman, J.J. (2017). All-Resolutions Inference for Brain Imaging. doi:10.1101/226126

[20] Sarkar, S. (2008). On the Simes inequality and its generalization. *IMS Collections Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen 1*, 231–242.