SAHLGRENSKA ACADEMY

Characterization of discrepancies between manual and automatic segmentation to improve anatomical brain atlases

M.Sc. Thesis

Anna Sörensson

Essay/Thesis:	30 hp
Program and/or course:	Medical Physics
Level:	Second Cycle
Semester/year:	Autumn 2020
Supervisor:	Rolf A. Heckemann
Examiner:	Magnus Båth

Abstract

Essay/Thesis:	30 hp
Program and/or course:	Medical Physics
Level:	Second Cycle
Semester/year:	Autumn 2020
Supervisor:	Rolf A. Heckemann
Examiner:	Magnus Båth
Keyword:	Anatomical brain atlas, Image segmentation, Image registration

Purpose: To characterize discrepancies between expert manually segmented brain images from Hammers Atlas Database and automatically generated segmentations of the same images; to decide whether they can be attributed to flaws in the automatic segmentation or in the manual segmentation; and to determine general rules that enable these decisions.

Theory: Image segmentation plays an important role in clinical neuroscience and experimental medicine for extraction of information from medical images, and it is a fundamental image processing step in medical image analysis. Another important image processing step is image registration that enables quantitative comparison between datasets of different subjects by geometrically aligning one dataset with another. The scientific underpinning of the project is descriptive science combined with inductive reasoning.

Method: The study data consisted of 30 T1-weighted 3D MR images along with manual region label volumes from Hammers Atlas Database, and automatically MAPER-generated segmentations of the same images. The comparison of manual and automatic anatomical (semantic) segmentations involves quantitative and qualitative analyses. Image registration was performed with MIRTK to normalize all images into a common space. Discrepancies were then extracted using a custom-designed image analysis process by the program Convert3D.

Result: The work has resulted in a model that enables extraction of discrepancies between manual and automatic segmentation into an individual component for quantitative characterization on a per-label basis. A total of 706 465 surface discrepancies were labelled while 1009 holes were found in both manual and automatic segmentations. Probability maps of the discrepancies have been created and can be used as a basis for determining the probability that certain discrepant voxels have been segmented correctly or not.

Conclusion: The study yielded insights into how differences between manual and automatic segmentations arise, and how these can be used to develop an improved segmentation that incorporates information from both models.

Sammanfattning

Bildsegmentering är en viktig del inom klinisk neurovetenskap och experimentell medicin vid insamling av information från medicinska bilder, och är ett viktigt bildbehandlingssteg inom medicinsk bildanalys. Det är därför viktigt med korrekt och noggrann segmentering men också att det finns etablerade metoder för att kunna undersöka och jämföra segmenteringsbilder. En annan viktig funktion inom medicinsk bildanalys är bildregistrering som möjliggör kvantitativ jämförelse mellan datamängder av olika sorters bilder. Processen bygger på att geometriskt anpassa en datamängd med en annan. Den vetenskapliga grunden för projektet är beskrivande vetenskap kombinerad med induktivt resonemang. Syftet med projektet var att karakterisera avvikelser mellan manuellt segmenterade hjärnbilder från Hammers Atlas Databas och automatiskt genererade segmenteringar av samma bilder för att avgöra om de kan tillskrivas som ett fel i den automatiska eller manuella segmenteringen, med målet att dra slutsatser om det finns allmänna regler som möjliggör dessa beslut.

I studien har 30 T1-viktade 3D MR-bilder med tillhörande manuell segmentering från Hammers Atlas Databas och automatiska MAPER-genererade segmenteringar på samma bilder använts. Jämförelse mellan manuella och automatiska anatomiska segmenteringar har involverat både kvantitativa och kvalitativa analyser. Bildregistrering utfördes för att normalisera alla bilder, och genomfördes med MIRTK. För att extrahera avvikelser mellan manuell och automatisk segmentering delades varje segmentering först upp i 95 binära ”regions”-bilder. Därefter multiplicerades varje automatiskt segmenterad binär regions-bild med 2 och adderades till motsvarande manuellt segmenterad binär regions-bild. Detta resulterade i en överlappad regionsbild per atlas. Bildbehandlingen utfördes i programmet Convert3D.

Studien gav en inblick i hur skillnader mellan manuella och automatiska segmenteringar uppstår och hur dessa kan användas för att nå en förbättrad segmentering som innehåller information från båda modellerna. Sammanfattningsvis har arbetet resulterat i en modell som möjliggör att avvikelser mellan manuell och automatisk segmentering kan definieras som individuella komponenter, och karakteriseras kvantitativt på en regionnivå. Sannolikhetskartor över avvikelserna har skapats och kan användas vid bestämning av sannolikheten för att en viss avvikande voxel har segmenterats korrekt eller inte.

Table of content

1. Introduction

Information from anatomical brain atlases enables, among other applications, image segmentation, pathology discovery, and identification of structure-functional relationships. Segmenting brain structures is instrumental for extraction of information from brain images and it is a fundamental image processing step in neuroimage analysis. It plays an important role in clinical neuroscience and experimental medicine by providing anatomical reference information for various uses (1, 2).

Several studies have been conducted with the purpose of creating protocols to enable anatomical labelling of the human brain. There have also been several attempts to improve previously generated protocols for more accurate results, as well as an expansion of the number of regions possible for segmentation. Manual segmentation demands a lot of skill and patience from the expert analyst and is time-consuming. With medical imaging technology evolving, the number and information content of medical images expanded to a point where the workload of expert visual analysis have become unsustainable, leading to a need for automatization. Several algorithms have been developed to enable automatic segmentation, resulting in faster segmentation and reducing the need of human input. In addition to this, several studies have emerged regarding validation and improvement of automatic segmentation where manual segmentation has been referred to as the golden standard surrogate of the ground truth, which is typically unavailable for *in vivo* images. But segmentation experts can make mistakes, and in the process of following atlas segmentation, following protocols, different interpretations or simple misinterpretations can occur.

In this project, manual and automatic segmentations are compared to characterize discrepancies between the two segmentation methods. This will be groundwork for further work to establish a typology of discrepancies with the view to determine their cause or other ways to determine if a discrepancy is being wrongly or correctly segmented. With a typology of this kind, it should be possible to establish rules for deciding whether individual discrepancies correspond to misclassifications in the manual, the automatic, or both segmentations.

1.1 Segmentation evaluation methods and metrics

The key aspect to image segmentation is the accuracy and precision with which structures can be detected and segmented. The quality of segmentation methods is of great importance in medical image analysis due to its strong impact on the outcomes. For example, in surgical planning this can affect the detection and monitoring of tumour progress and have a direct impact on the results. Evaluation of segmentation methods is therefore important and the different evaluation approaches have been widely studied. There are various evaluation metrics for comparing quality in medical image segmentation. Depending on the parameter the metric evaluates, they can be based on overlap, volume, probability etc. The point of an evaluation metric is to give an indication of the errors in a segmentation, based on its data and the segmentation task. Some of the most common evaluation metrics for comparing different segmentation methods are the Dice coefficient, Jaccard index, volumetric similarity and mutual information which is overlap-, volume- and information theoretic based (3). Usually these metrics are applicable when one wants to evaluate a segmentation against another one that serves as the standard reference segmentation, e.g. a manually segmented reference image. The evaluation approach of this kind is called supervised segmentation evaluation. The opposite evaluation approach is called unsupervised segmentation evaluation and the key advantage of this approach is that it does not require a reference segmentation image. This type of approach enables evaluation of any segmented image but also the unique potential for self-tuning. The potential advantage the supervised methods have over unsupervised methods is that "the direct comparison between a segmented image and a reference image is believed to provide a finer resolution of evaluation" (Zhang et al. 2008, p. 261) (4). Even though the unsupervised method seems more beneficial, the supervised is more commonly used because it is easier to apply, leading to the importance of finding new unsupervised methods. In this project, the work conducted will be a foundation in an approach to create a new unsupervised method to study segmentation images.

1.2 Image registration

In addition to image segmentation, image registration is another important processing step in image analysis that enables quantitative comparison between datasets of different subjects by geometrically aligning one dataset with another. It also provides the possibility to combine information between images from different modalities, collected at different times and/or by various detectors for comparison. Working with brain atlases, whether it is atlas construction or studying structure and functional organization of the brain, image registration is a requirement (5).

There are numerous approaches for image registration, but the principle is that two or more images are spatially transformed into one coordinate system by an algorithm. Upon registration, one image is chosen as the reference image and the other one/-es is referred to as the source or floating image/-es. The reference image is kept untouched while a geometric transformation is applied on the source image/-es to align it with the reference image (6). Commonly, registration follows a multi-level hierarchical model where the alignment between the images improves in successive steps by applying geometric transformations with increasing numbers of degrees of freedom. Accordingly, improvements go from a coarse to a fine detail level while the output generated at each step is used as the starting point for the next.

The first and simplest transformation is rigid which includes the geometric operators; translation and rotation. In 3D, each operator applies in all direction (x, y, z), giving rise to three parameters or degrees of freedom each. Consequently, rigid transformation can be defined by six parameters. Affine transformation includes translation, rotation, scaling and shear, and is defined by 12 parameters while non-rigid transformation includes translation, rotation and uniform scaling and is defined by 9 parameters. Rigid and affine transformation are considered global transformations, which means that global distortions between the images will be corrected when the transformations are applied in a registration. If registration includes a non-rigid transformation, which is considered a local transformation, local deformation will be corrected. Selection of geometric transformation method depends on the nature of the registration data (7, 8), e.g. if the registration is inter- or intra-subject, multimodal etcetera. For example, a registration of images from different subjects (inter-subject) requires additional degrees of freedom to account for all the possible deformations between the images to be aligned.

When working with image registration there are other factors than selection of geometric transformation that must be considered. Two other factors are similarity measures and interpolator. The similarity measures can be classified into intensity- and feature-based methods. There are similarity measures that can be included in both classes, depending on the features used. The first operates with information based on intensity differences, intensity cross-correlation and information theory, while the second operates with structures like regions, lines and points in the images. For intensity-based measures, sum of squared differences (SSD), correlation ratio, mutual information (MI) and their derived measures, can be used. SSD and MI can also be used in feature-based methods. There are numerous derived quantities, such as normalized MI and correlation ratio. The aim of interpolation is to estimate the intensity of a point after transformation to a new position. There are many different interpolation algorithms available. Some common ones are based on nearest neighbour, linear interpolation, cubic B-spline and windowed sinc interpolations (7, 8). Selection of the factors and their alternatives mentioned above also depends on the type of registration data. For example, registration with inter-subjects require correlation ratio or MI (or some of their derived measures) while for intra-subjects, SSD is more suitable. For example, if the registration is performed on MR intensity images with continuous scale, linear interpolation can be a suitable choice. However, for label images that use a categorical scale, intermediate values between integers have no defined meaning. In this case, interpolation with nearest neighbour is more adequate.

1.3 Aim

The purpose of this project was to to characterize discrepancies between expert manually segmented brain images from Hammers Atlas Database and automatically generated segmentations of the same images; to decide whether they can be attributed to flaws in the automatic segmentation or in the manual segmentation; and to determine general rules that enable these decisions. The idea is that, based on the results of these investigations, it will be possible to create new improved brain atlases, which may potentially lead to improvements of the Hammers Atlas Database but also enable clearer quality assessment of automatic brain image segmentation approaches. The work will yield a software code base for implementing open-source software tools that detect (and optionally fixing) common flaws in manual as well as automatic segmentations.

2. Material and methods

The following section gives information of the material used in the project and presents methods employed.

2.1 Hammers Atlas Database

2.1.1 Background

The Hammers Atlas Database is a publicly shared resource that consists of 30 T1-weighted 3D magnetic resonance (MR) brain images with corresponding manually generated anatomical label sets, maximum probability atlas, regional probabilistic maps, and participant demographics. The atlases are segmented into 95 regions each and are provided under academic licence at www.brain-development.org (9). The development of the Hammers Atlas Database started in 2000, when Alexander Hammers and his group studied 20 healthy adult volunteers (10 women) with a median age of 31 years. Images were obtained with a 1.5 Tesla GE Signa Echospeed scanner at the UK National Society for Epilepsy, using inversion recovery prepared fast spoiled gradient recall sequences to create T1-weighted 3D volumes of the whole brain. Initially, an anatomical labelling protocol for 49 regions in the human brain was developed (10). Later, ten additional participants were recruited and the protocol extended to 83 region (11). The work continued and six regions were added in Wild HM et al. 2017 (12), followed by six more in Faillenot I et al. 2017 (13), leading to a total of 95 regions segmented in 30 MR brain atlases. Further expansion is ongoing. The data collection from the volunteer participants took place with ethical permission, so by adhering to the terms of the above mentioned academic licence, all ethical obligations in connection with the present study are fulfilled.

2.1.2 Use in the project

For the present study, all 30 available T1-weighted 3D MR images were used, along with the corresponding manual region label volumes. The region labels were supplied as images spatially correlating with the MR image, where each voxel was labelled with a value from 1 to 95 (corresponding to the 95 regions) or 0 (for background, i.e. non-brain portions of the image or brain regions not included in the protocol).

2.2 Automatic image segmentation

MAPER (multi-atlas propagation with enhanced registration) is a method for anatomically segmenting MR images of the human brain. MAPER is written and maintained by the supervisor of this project (2). The method is based on previous work on multi-atlas based segmentation (1). MAPER allows automatic delineation of regions on newly acquired images or already existing ones using the knowledge embedded in already existing atlas databases. This method was developed by using the Hammers Atlas database, but can be applied with other manually segmented atlases (14). MAPER segmentations of all 30 T1-weighted images from the Hammers Atlas Database were available for use in the present project.

2.3 Image normalization to a common space

All 90 image data sets (30 MR, 30 manual segmentations, 30 automatic segmentations) were normalized to a common space by geometric transformation to a common template. Image registration was carried out with MIRTK (medical image registration toolkit), a research-focused image processing toolkit for image data processing and analysis (15). For the normalization target, MNI152 was used; a template created by averaging 152 T1-weighted images of healthy adults (15, 16). Each MR image was paired with the MNI152 atlas. The images were aligned by maximizing normalized mutual information (NMI) as the similarity measure and the geometric transformations rigid-, affine- and non-rigid free-form deformation (FFD) based on B-splines. The FFD model, based on B-splines, is a local transformation which models 3D deformable objects. Cited from Rueckert et al. 1999, p.713 “The basic idea of FFD’s is to deform an object by manipulating an underlying mesh of control points” (8). By performing registration with all the steps of geometric transformation, both global and local details were corrected for. The result from the registration process was 30 matrix transforms. These were later applied to the corresponding segmented images by using nearest-neighbour interpolation, separately for the manual and automatic segmentations.

2.4 Image processing

The extraction of information about discrepancies between the manual segmentations and the corresponding automatically generated segmentations was performed through the software application ITK-SNAP (version 3.6.0, April 1 2017) and the complementary application Convert3D (c3d) in the common space. The software application can be found in the link; http://www.itksnap.org and provides, among other things, image visualization and navigation (17). The application c3d, found in the following link; https://sourceforge.net/p/c3d/git/ci/master/tree/doc/c3d.md, is a command-line image processing tool that offers complementary features to ITK-SNAP, e.g. various tools specialized for multilabel images and multicomponent images (18).

Convert3D offers a function called "holefill", which enables localization and filling of holes in the labels. This process was done separately for the manual and the automatic segmentations for each atlas and labels. Subsequent to the find-and-fill hole process, the "holefilled" label images were subtracted from the corresponding intact image labels, giving output images consisting exclusively of the holes for each respective label, atlas and segmentation method. The purpose of this process was to identify discrepancies that were not on the surface of a label.

To detect surface discrepancies, the manual and automatic generated atlases were holefilled and split into 95 binary label images each, one image for each region. Thereafter, each automatic segmented binary label image was multiplied by 2 and added to the corresponding manually segmented binary label image for each atlas, resulting in 95 "overlaid" label image per atlas. The overlay image consisted of voxels with the values 0, 1, 2 and 3 (see Figure 1). Voxels with the value 0 and 3 represents the background and the foreground agreement (overlap area) respectively. Areas with voxel value 1 represent voxels that had been identified by the manual segmentation but not by the automatic (error type f1), and the other way around for voxels with value of 2 (error type f2). The next step was to separate the error types from each other and extract each discrepancy. Voxels with the value 3 were set to 0, resulting in label images containing only the values 0, 1, and 2. The label images were thereafter split into two binary images, one image with error type f1-voxels and one second image with the error type f2-voxels. After separating the error types, the different connected components were separated by using the c3d built-in command "comp". By once again splitting the images, each connected component (discrepancy) was extracted, see code in Appendix 1. To summarize, the whole process generated binary discrepancy labels for each atlas, each label and each type (manual and automatic), producing a binary image for every individual surface discrepancy.

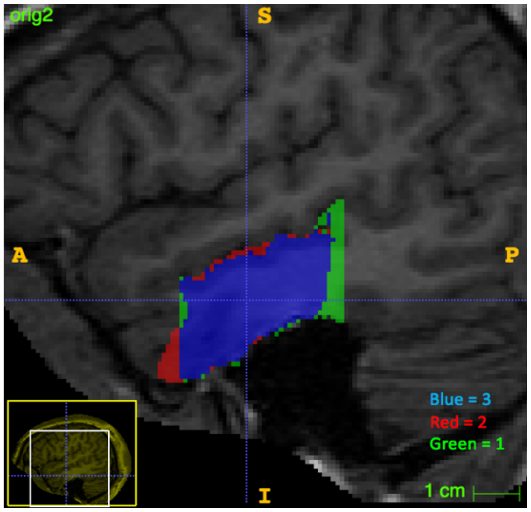


Figure 1: Region 13 (right middle and inferior temporal gyrus) in a sagittal section. The red area represents voxels labelled by only the manual segmentation (voxel value 1), green area represents voxels labelled only by the automatic segmentation (voxel value 2) and the blue area is the foreground agreement (voxel value 3). The background agreement (voxel value 0) is not highlighted.

2.5 Probability maps

Probability maps of the discrepancies were created to find frequency relationships across the 30 individual atlases. All discrepancies with the same error type and from the same region across the 30 atlases were added together, resulting in 95×2 discrepancy summation maps that consist of voxels with values between 1 and 30. To acquire the probability that a specific voxel occurs, the summation maps were divided by the total number of images added (30 in our case). This gave an overview of which discrepant voxels occur most often. A probability map consisting of discrepancies due to automatic segmentation can give an indication of where the systematic errors are located. Furthermore, it can also indicate which discrepant voxels should be included in the associated label due to the fact that most segmentations have considered these exact voxels to be in the referred region. A high voxel value corresponds to high probability and means that the voxel was discrepant in several atlases and the other way around for low voxel value.

2.6 Data collection and analysis

Using c3d, the number of connected components for each label pair and error type was determined. Due to a limitation in c3d, only the 254 largest components per pair and error type were further analysed. The following characteristics were determined for each individual discrepancy component image: atlas number, label number, error type, voxel count, centroid coordinates, and extents in the x, y, and z directions. The resulting data were loaded into R Version 4.0.3 for descriptive analysis (https://www.r-project.org). Relative volume was calculated by dividing the discrepancy volume with the union of the corresponding manual label volume and automatic label volume.

2.7 Visual comparison

From the quantitative analysis, a set of regions was chosen for a qualitative comparison to characterize the discrepancies. For each region, the five largest surface discrepancies were selected for the visual comparison. The regions chosen were left and right occipital lobe (label 22 and 23). The underlying argument for selecting these regions came from the processed data collected from image processing and probability maps. The chosen discrepancies are listed in Appendix 2.

Initial investigations to qualitatively characterize discrepancies between manual segmentations from the Hammers' atlas database and automatically generated segmentations on the same images by visual comparison have been reported (1). In these investigations the manual segmentation was used as the reference frame. The discrepancies were classified by error type based on its appearance, creating a typology for qualitatively characterized discrepancies. In the study, five different error types were defined (cf. table 1) that indicate the shape of the discrepancy, how they were related to the label volume but also a way to conclude which label was correct. The first error type, random error (RND) is described as small discrepancies due to interpolation while the second error type, greedy/shy labeling (GSL) is defined as "error that systematically places the label boundary beyond or short of the reference label but preserves its shape" (Heckemann et al. 2006, p. 119). A discrepancy of this error type would be a thin layer due to automatic segmentation on the foreground agreement, which places the label boundary beyond the reference label. Label propagation failures (LPF) are discrepancies due to the automatic segmentation that are composed of connected voxels assigned to a structure in error. Manual segmentation failures (MSF), however, are discrepancies due to the manual segmentation that were found in retrospect questionable. The last error type is planar boundary error (PBE) which occurs when a knowledge-based boundary is displaced (1).

Visual analysis of the discrepancies within labels was also carried out with the aim of investigating whether holes existed that should not be considered flaws. A test sample with the holes of largest volume size and with both error types was chosen for visualization. Additionally, 5 discrepancies due to manual segmentation with much smaller volumes were also examined. The discrepancies chosen for the visualization are listed in Appendix 3.

Table 1: An overview of the error type. Reproduced with permission from Heckemann et al. 2006 (1).

Summary of error types

Abbreviated error type	Error type	Description
RND	Random error	Individual discrepant voxels resulting from interpolation
GSL	Greedy/shy labeling	Connected areas of discrepancy, following the structure outline
LPF	Label propagation failure	Misassignment of a group of connected voxels to a structure
MSF	Manual segmentation failure	Discrepancies due to mistakes in the expert segmentation
PBE	Planar boundary error	Misplacement of a knowledge-based boundary

3. Results

3.1 Surface discrepancies

Across 30 atlases with 95 regions, all regions (2850) showed discrepancies between the automatic and manual segmentations. The discrepancies were of both error type f1 (segmented by the manual method but not by the automatic) and type f2 (segmented by the automatic method but not by the manual). A total of 706 465 surface discrepancies were labelled. The most discrepancies for an error type, 1134, was found in label 18 (left cerebellum) of atlas 26 with error type f1 (see Figure 3), while the minimum number of discrepancies for an error type f2, was found in label 79 (right frontal lobe subcallosal area) of atlas 7 with error type f2. Notice that figure 3 only shows the maximum total number of discrepancies found in a label, regardless of error type. The total number of discrepancies in a label for all atlases are not presented here. In summary, the largest and second largest number of discrepancies were found in label 17 and 18, left and right cerebellum, for all atlases except for atlas 3, 16, 28, 29 and 30. The union of the manual and automatic label volume for these labels was found to be the largest for almost all atlases as well. Figure 4 shows the frequency of the discrepancies found in each label. The data in the histograms consisted only of discrepancies up to 254, due to the limitation of the "split"-command in c3d. The number of discrepancies for each label varies, but most discrepancies were found in labels with larger volumes. A similar relationship was found between the discrepancy volume and the label volume, where larger discrepancy volumes occur more frequently in labels with larger volumes.

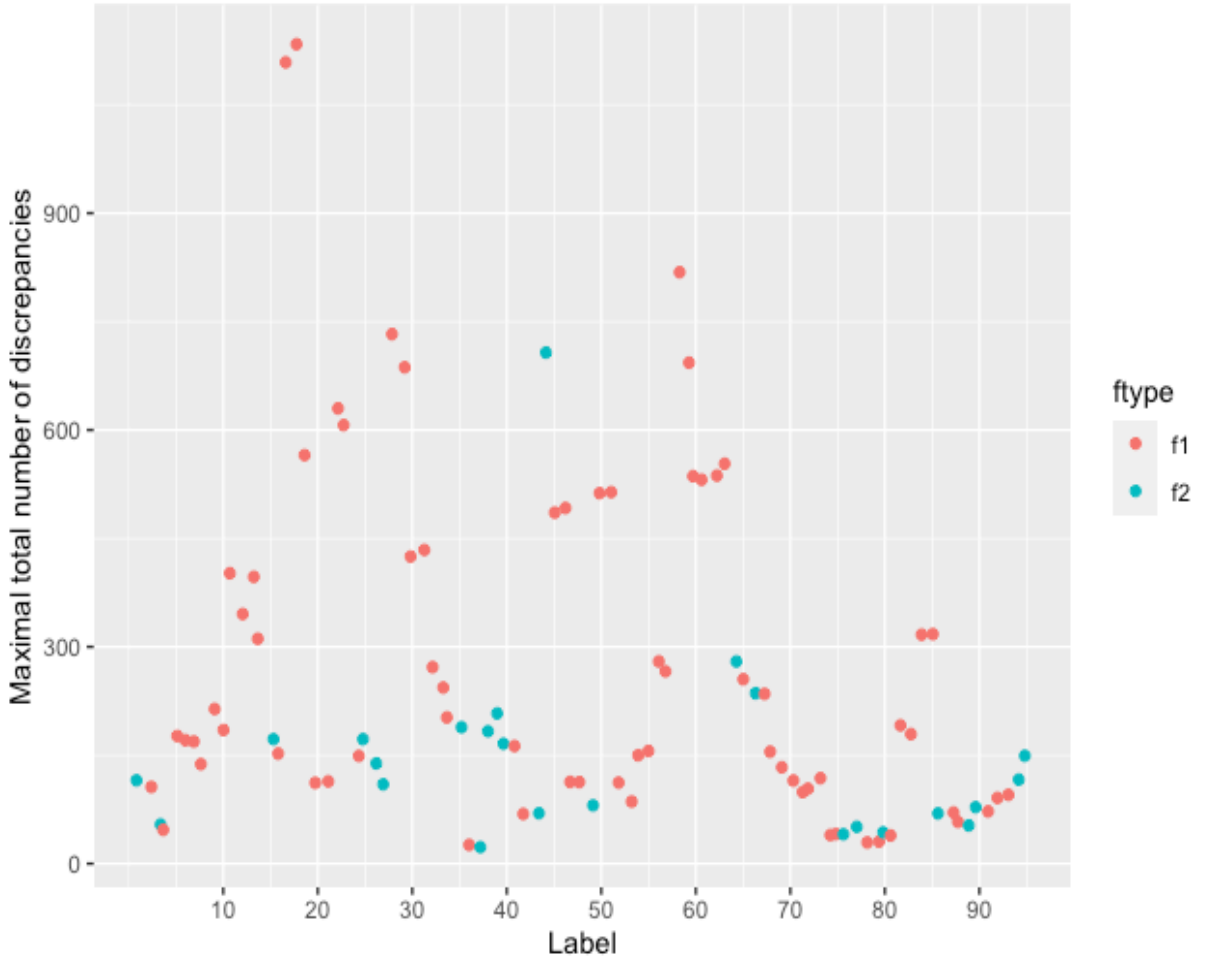


Figure 3: An overview of the maximal total number of discrepancies found in each label. The points differ in colour which present the error type of the discrepancies. The error type f1 represents by colour red and error type 2 represent by colour blue.

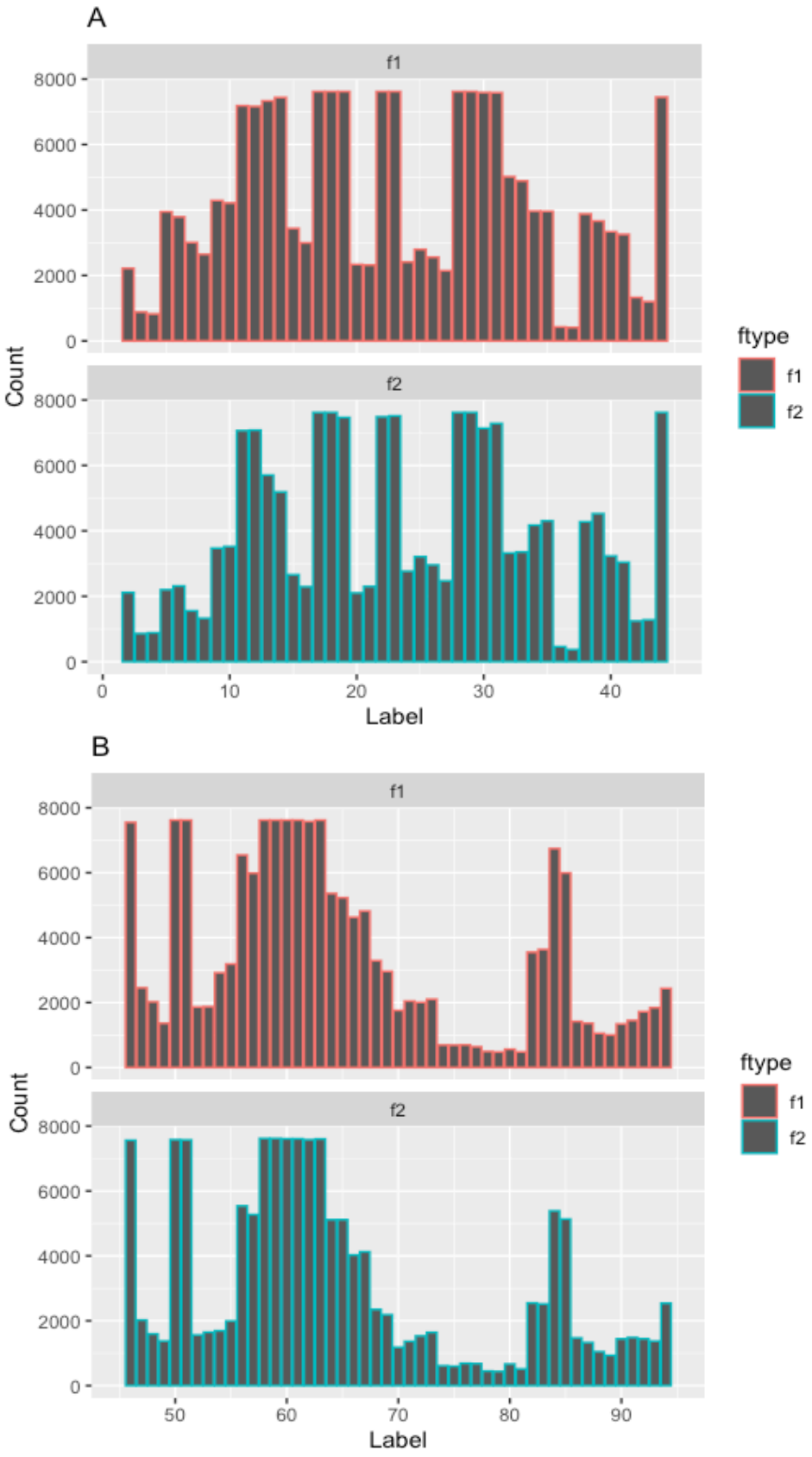


Figure 4: Frequency histograms of surface discrepancies in each label, where A presents the number of holes in label 1-45 and B for label 45-95. The holes due to manual (f1) and automatic (f2) segmentation are shown in separately histograms.

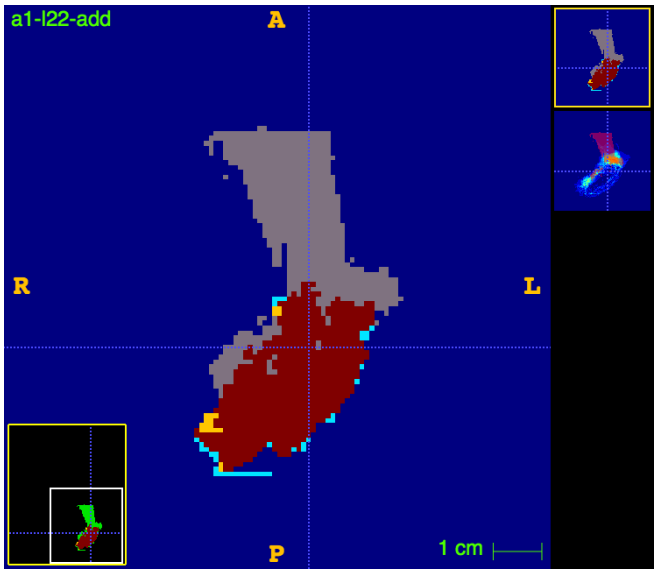
3.1.1 Visual analysis of surface discrepancies

The shape of the discrepancies that were visually studied can be described by the structure definitions in the error types GSL and LPF. Most discrepancies were of relatively large volume and their shape could be categorized under LPF. Some of them, even though of large size, were formed like a thin layer on the foreground agreement (volumes with voxel value 3) that does not affect the structures outline. Thin layer in this sense are connected voxels, forming a line in one of the dimensions but are a few voxels wide in the rest of the dimensions' range. These kinds of discrepancy can be categorized under the structure definition in GSL.

One of the discrepancies (a1-l22-f1-c1) that could be categorized under LPF was the largest connected component found (see Figure 5A), which was due to the manual segmentation and located in the left occipital lobe (label 22), atlas 1. This discrepancy was located in a way that form a large extended region relative to the foreground agreement. A similar discrepancy (a1-l23-f1-c1) as a1-l22-f1-c1 was found in the corresponding label on the opposite side of the brain, left occipital lobe (label 23). The shape was distinctly alike and located similarly but mirrored (see Figure 5B). One observed discrepancy (a5-l22-f1-c4) that was hard to categorize was shaped as a relative thin layer, but was not located on the foreground agreement. It was rather a thin slice into the foreground agreement and surrounded of foreground agreement voxels.

In figure 6 and 7, the summation maps of label 22 and 23 are shown with and without the discrepancies a1-l22-f1-c1 and a1-l23-f1-c1 as an overlay. Analysing the discrepancies as an overlay on the corresponding summation map showed that many of the voxels in these discrepancies (a1-l22-f1-c1 and a1-l23-f1-c1) has not been segmented as these labels (22 and 23) in other atlases due to the fact that the colour of these voxels is a shade of the blue representing the background.

A



B

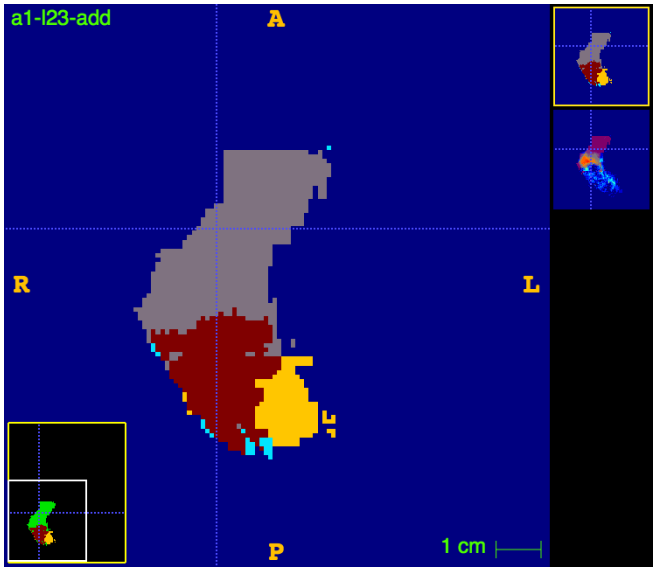
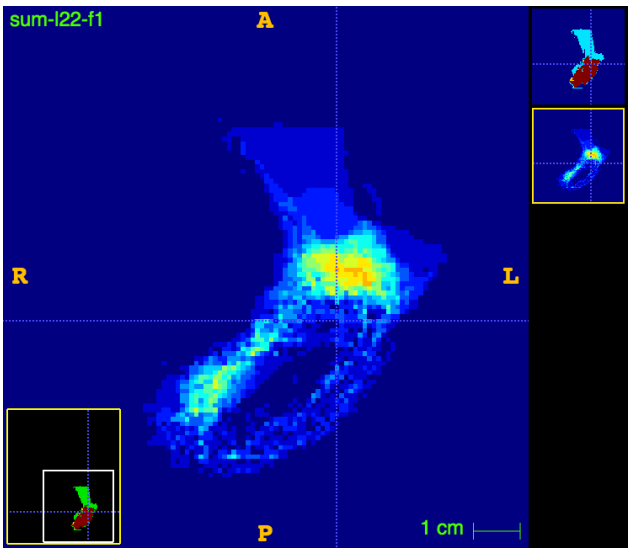


Figure 5: Left and right occipital lobe (label 22 and 23 respective) where the discrepancy a1-l22-f1-c1 is outlined as the grey area in A, and the discrepancy a1-l23-f1c1 as the grey area in B. The images are in a transversal section.

A



B

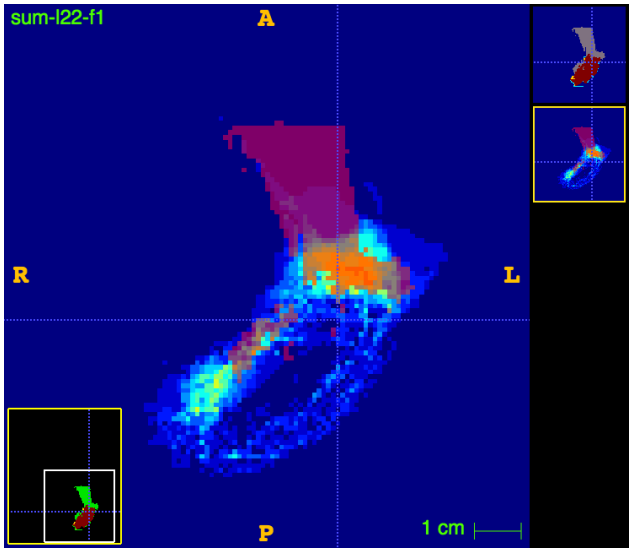
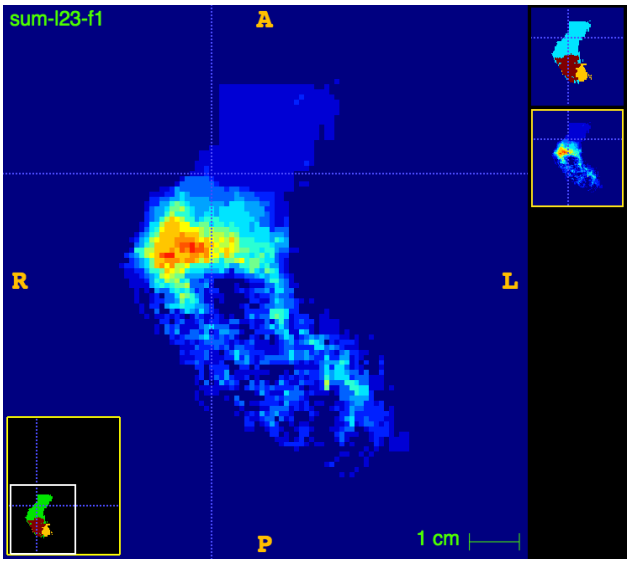


Figure 6: Summation map of left occipital lobe (label 22) for error type f1, where the discrepancy a1-l22-f1-c1 is outlined as the purple overlay area in B. The summation maps include voxels with values between 0 and 30, where voxel value 0 is background and a higher number presents a more frequent occurring voxel. The colour scale in the figure 6 and 7 goes from blue to red, where blue and red represent lower versus higher voxel value respectively. The images are in a transversal section.

A



B

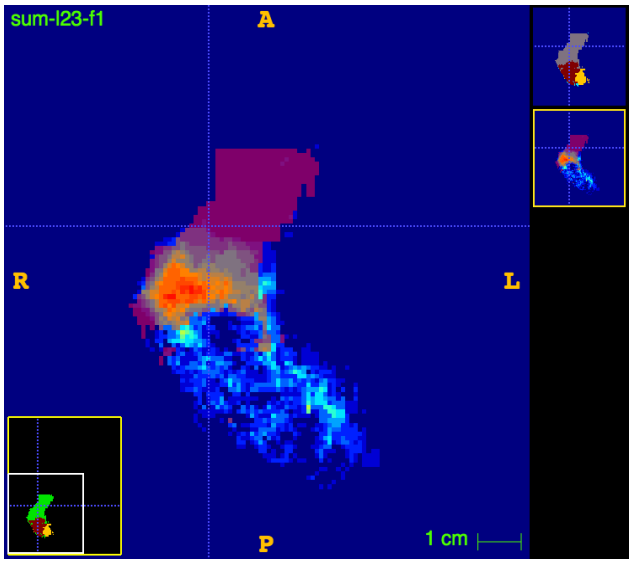


Figure 7: Summation map of right occipital lobe (label 23) for error type f1, where the discrepancy a1-l23-f1-c1 is outlined as the purple overlay area in B. The summation maps include voxels with values between 0 and 30, where voxel value 0 is background and a higher number presents a more frequent occurring voxel. The colour scale in the figure 6 and 7 goes from blue to red, where blue and red represent lower versus higher voxel value respectively. The images are in a transversal section.

3.2 Label holes

From the "holefill" process, 1009 holes were found in the manual and automatic segmentation in total. Most of the holes occurred in the automatic segmentations and the largest number of holes occurred in label 23 (right occipital lobe) (see Figure 8 and 9). In most of the labels where holes were identified, holes occurred for both segmentations methods even though they occur more frequently in the automatic segmentation. The number of holes for the manual segmentations was, however, substantially larger in label 17 and 18. In Figure 10, the discrepancies' volumes are shown. Most of the manual holes were significantly bigger than the automatic holes. Most of the holes in the automatic segmentation labels were of single voxels, whereas the manual holes had volumes up to 206 voxels. Ranking the holes after the volume size, the first 84 holes were due to the manual segmentation. The largest automatically segmented hole had a volume of 7 voxels.

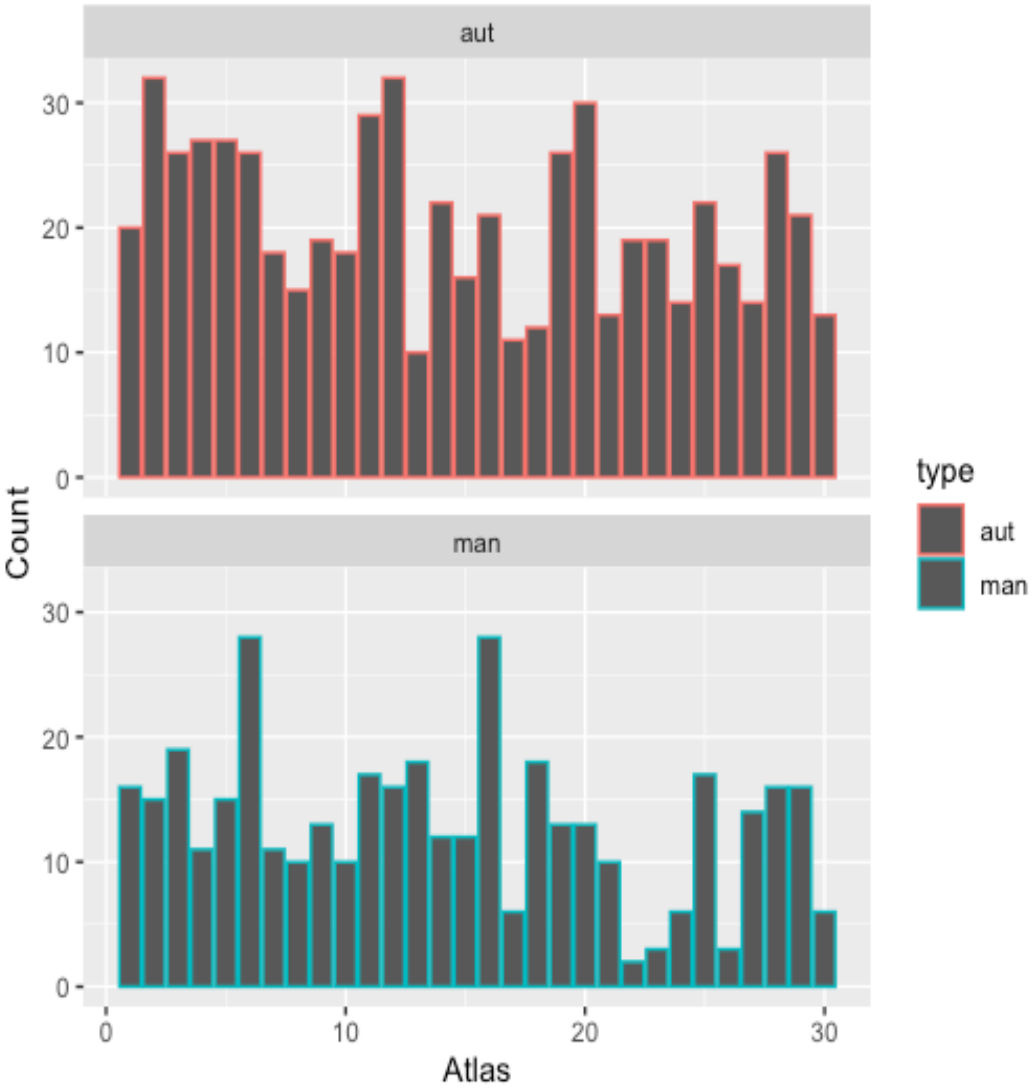


Figure 8: Frequency histogram of the holes across the 30 atlases, found in manual (blue piles) and automatic (red piles) segmentation separately.

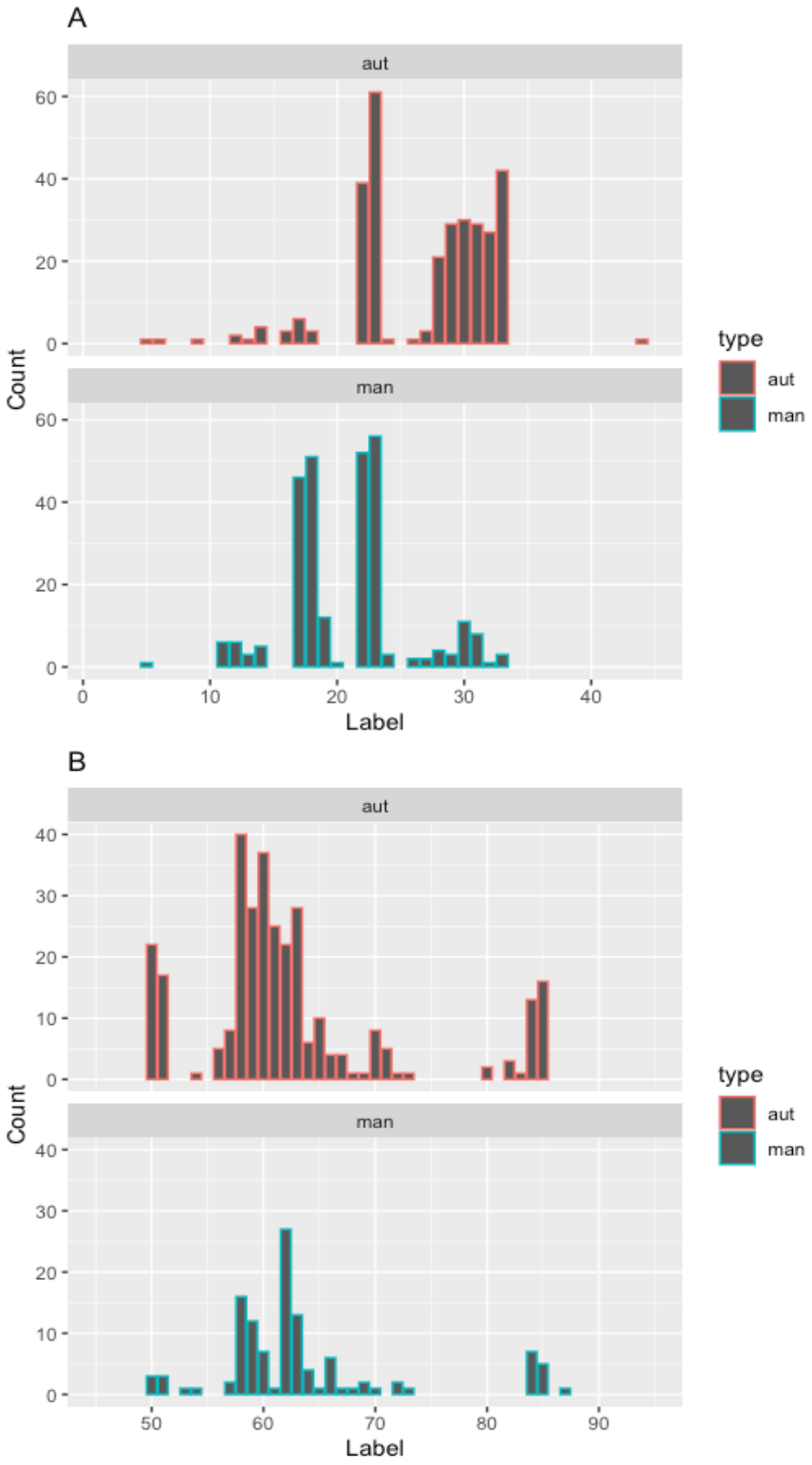


Figure 9: Frequency histograms of holes across the labels, where A presents the number of holes found for label 1-45 and B for label 45-95. The holes due to manual and automatic segmentation are shown in separate histograms as blue and red piles respectively.

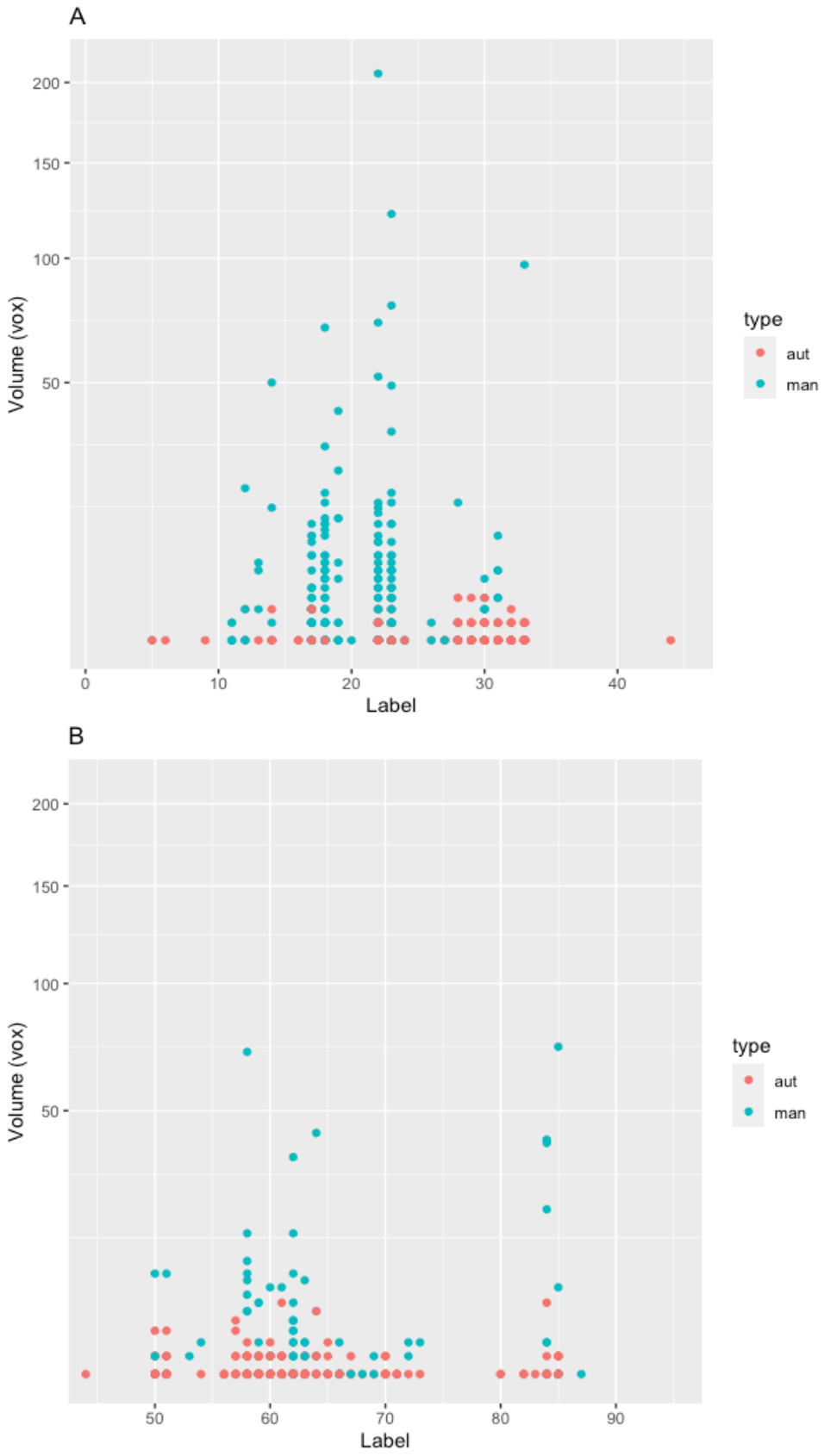


Figure 10: An overview of the hole volumes across the 95 labels for manual and automatic segmentation as red and blue points respectively. A) Hole volumes for label 1-45, and B) for label 45-95. The volume is presented as the numbers of voxels.

3.2.1 Visual analysis of label holes

In table 2, a summation of the visual analysis and personal comments of the chosen discrepancies are presented. All holes have been labelled with a "type" based on visualization in the MR images. Holes that are likely to be wrong have been labelled as "FALSE" while the holes that seem to be legitimate have been labelled as "TRUE". Some holes have been given the label "unclear", which means that it was unclear if they could be a flaw or not. In most of these unclear cases, the holes were either partly in low-intensity regions that might correspond to CSF or partly on brain tissue.

Table 2. The analysed holes named by the atlas, type error and label it was located in. "c1" refer to the first largest discrepancy found in the label, for that specific atlas. "Type" indicates whether the holes seem to be true, false or unclear (fussy) and the voxel value is the value given by the segmentation methods.

Name	Volume (voxels)	Type	Voxel value	Comments
a6-h-man-l18-c1	70	Unclear	0	Partly on CSF, partly on brain tissue
a2-h-man-l58-c1	71	FALSE	24	Close to neighbouring region 24
a2-h-man-l14-c1	50	TRUE	0	
a25-h-man-l33-c1	97	Unclear	23	Partly on CSF, partly on brain tissue
a25-h-man-l23-c1	122	TRUE	45	Seems like a legitimate hole from MR-image
a24-h-man-l22-c1	206	TRUE	46	Seems like a legitimate hole from MR-image
a1-h-man-l85-c1	73	TRUE	0	
a16-h-man-l22-c1	52	TRUE	0	
a11-h-man-l22-c1	72	TRUE	0	
a10-h-man-l23-c1	79	Unclear	31	Partly on CSF, partly on brain tissue
a25-h-aut-l84-c1	7	Unclear	60	Close to CSF but not on, embedded in the region
a24-h-aut-l61-c1	7	TRUE	51&63	Seems like a legitimate hole from MR-image
a10-h-aut-l64-c1	6	Unclear	22	Unclear
a16-h-aut-l57-c1	5	TRUE	29	Seems like a legitimate hole from MR-image
a8-h-aut-l57-c1	4	TRUE	0	
a5-h-aut-l23-c1	4	Unclear	33	
a26-h-aut-l50-c1	4	TRUE	28	Seems like a legitimate hole from MR-image
a1-h-aut-l29-c1	4	FALSE	51	Embedded in the region
a17-h-aut-l51-c1	4	Unclear	29	Seems maybe to be on a CSF spot, unclear
a13-h-aut-l30-c1	4	FALSE	22	Very close to neighbouring region 22, surrounded by holes
a28-h-man-l51-c1	11	FALSE	59	Very close to neighbouring region 59
a28-h-man-l50-c1	11	Unclear	0	Seems to be partly on brain tissue and partly on CSF
a19-h-man-l22-c1	11	FALSE	30	Very close to neighbouring region 30
a15-h-man-l17-c1	11	TRUE	0	
a14-h-man-l22-c1	11	TRUE	0	

The visual analysis of the 10 largest holes in manual segmentations shows that most of them are not flaws. Most of these holes are located in regions where there are sulci with cerebrospinal fluid (CSF) going from the subarachnoid space into the brain, creating space with CSF within some regions and furthermore a hole when the surrounding areas are segmented as brain tissue. The expert has in these cases assigned a 0 (background) label to the voxels which form the holes, defining these voxels as part of a sulcus with CSF. In cases where a hole seems wrong, or it is unclear if a hole is legitimate, the expert has segmented these voxels as a voxel value of the neighbouring region. There are also holes in ventricle regions. These voxels have also been segmented as a voxel value of neighbouring region. One could argue that these kinds of holes are true holes but then the voxels forming the holes should not be labelled as a neighbouring region. Figure 11 shows a case where a hole corresponds to a deep sulcus. The hole shown in Figure 12, corresponds to a disconnected part of a ventricle that appears as a hole in label 22 (Left occipital lobe). Figure 13, shows a hole in the brain tissue.

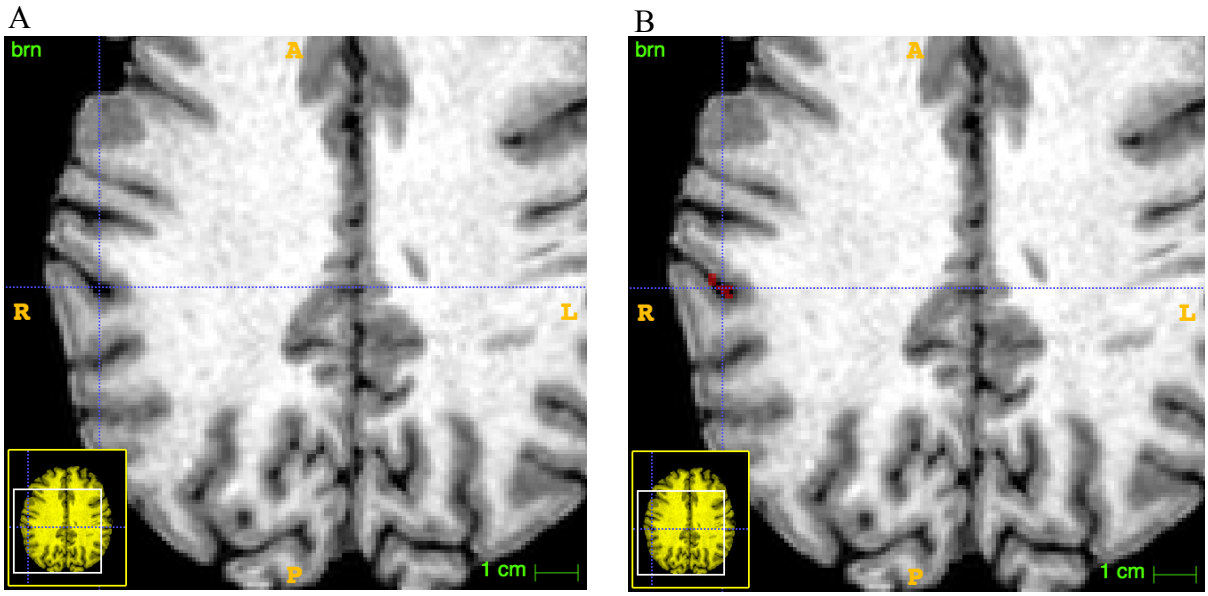


Figure 11: MR-image of right supramarginal gyrus (label 85) in atlas 1 A) without, and B) with the label hole a1-h-man-l85-c1 outlined in red in a transversal section.

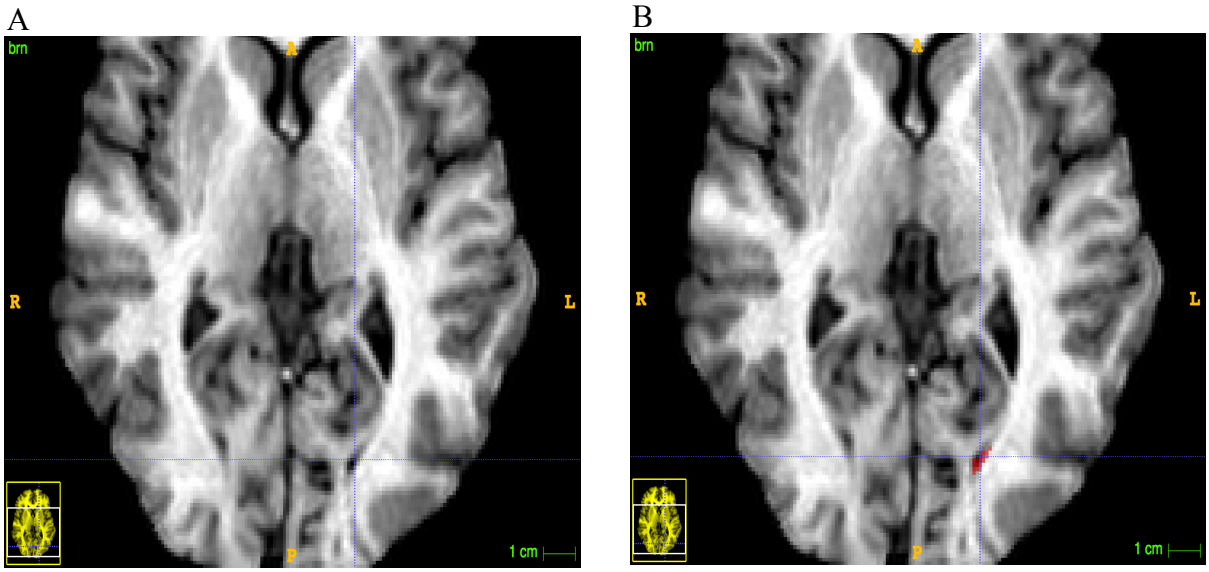


Figure 12: MR-image of left occipital lobe (label 22) in atlas 24 A) without, and B) with the label hole a24-h-man-l22-c1 outlined in red in transversal section.

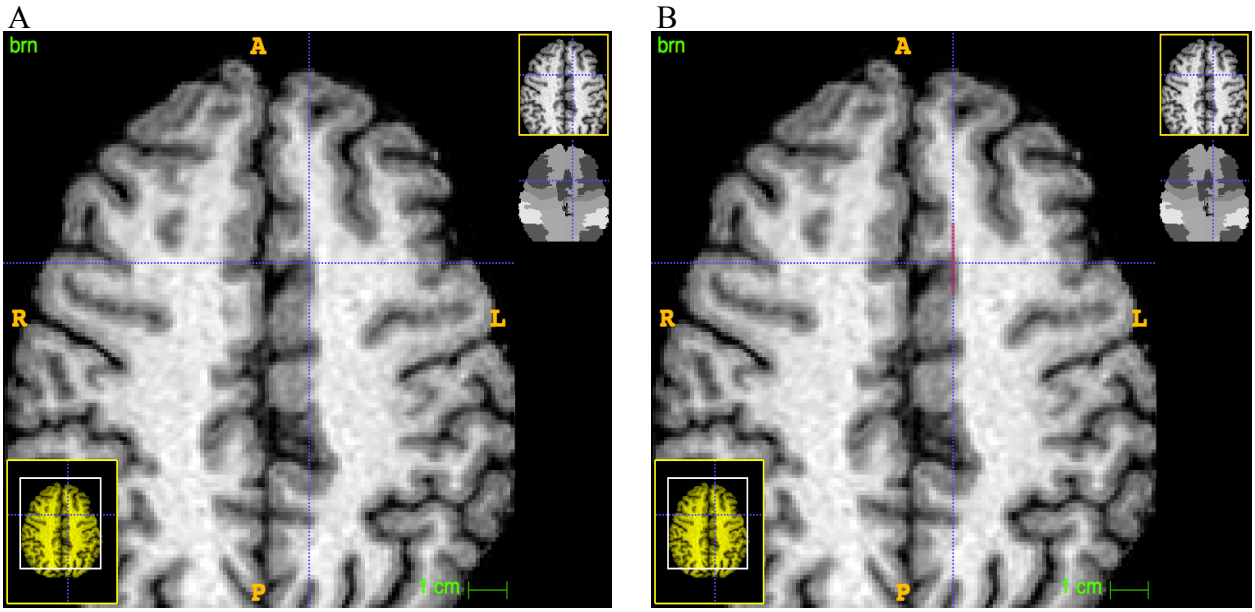


Figure 13: MR-image of left superior frontal gyrus (label 58) in atlas 2 A) without, and B) with the label hole a2-h-man-l58-c1 outlined in red in a transversal section.

Of the five smaller holes due to manual segmentation, holes were found to be both true and false. The holes that seem to be wrong were located in the brain tissue and consisted of voxels with the value of a neighbouring region, while voxels in true holes have voxel value 0. For the 10 largest holes in automatic segmentation labels, it was hard to determine whether they were true or false. Some of the holes were potentially true, because they corresponded to CSF, but viewing the automatically produced brain atlas one could see that the voxels forming the holes have been segmented as a neighbouring region. In most of the cases where the hole was determined to be wrong, the hole was embedded within a region but there were some cases that deviated from this trait. Analysing a hole that seems to be false as an overlay on the corresponding segmentation image, one could see that the hole was very close to the border of neighbouring region whose value the hole has been assigned. One of these holes (a14-h-aut-l30-c1) were also close to other smaller holes with same labelled voxel value, see Figure 14.

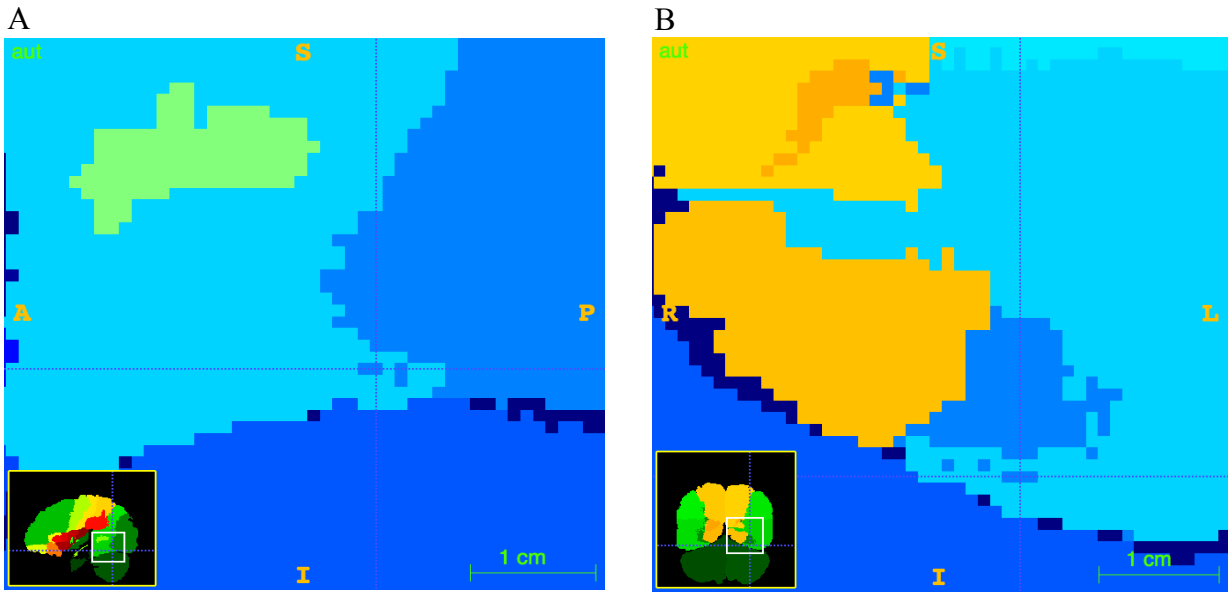


Figure 14: Automatic segmentation image for atlas 14 in A) sagittal, and B) coronal section showing areas around label 30 and its neighbouring labels. The cross point out the hole a14-h-aut-l30-c1.

3.3 Probability maps

For an easier understanding, the results from the summation maps (sum-maps) will be presented as integer values instead of floating-point numbers from the probability maps. This means that the summation maps have not been normalized and the probability will be given by dividing the result values with 30.

The maximum voxel value across the sum-maps was 19, corresponding to a probability of 63% (19/30). This occurred in four sum-maps corresponding to the following regions; left posterior temporal lobe, left middle frontal lobe, left supermarginal gyrus and right superior parietal gyrus. All four of these sum-maps consisted of discrepancies due to manual segmentation (error type f1). The smallest value of the maximum voxel value was 10, corresponding to 33 % and occurred in the sum-maps of the following regions; right lateral ventricle temporal, left insula posterior short gyrus and right ventricle excluding temporal horn, (see Figure 15). The discrepancies in these sum-maps were also due to manual segmentation (see Table 3). In figure 15, an overview of the sum-maps' voxel value range can be seen. Most sum-maps have a maximum voxel value between 14-16 (46-53%). The mean maximum voxel value is 49 % (14.6/30) which seems to correspond with the frequency histogram presented in Figure 10.

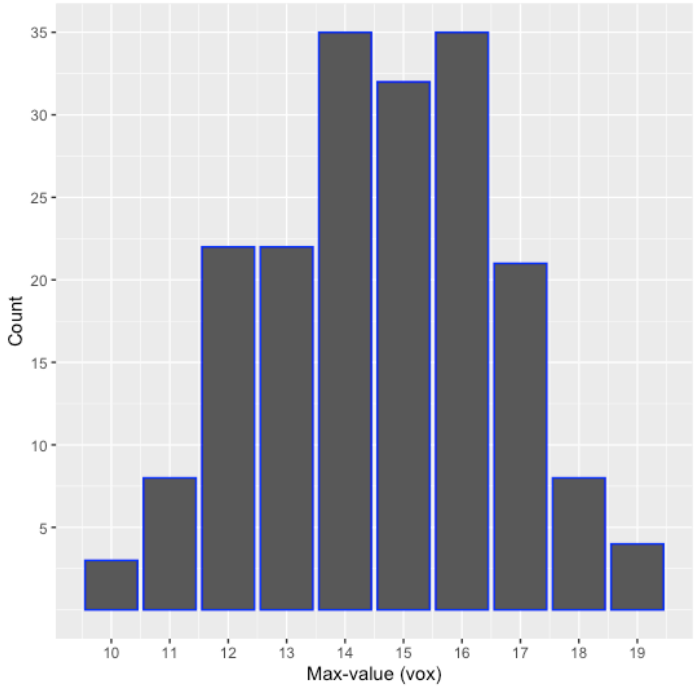


Figure 15: The frequency of the summation maps' maximum voxel value

Table 3: The summation maps with highest and lowest maximal voxel value and their summation voxel value, volume and mean voxel value. Label-ftype provides label and error type information. The error type f1 corresponds to a discrepancy due to manual segmentation. The probability is given by dividing max-vox-value by 30.

Label-ftype	Max-vox-value	Sum-vox-value	Volume (vox)	Mean-vox-value
Sum-l30-f1	19	290 559	69 852	4,159638
Sum-l84-f1	19	177 175	48 698	3,638240
Sum-l28-f1	19	327 222	90 668	3,609013
Sum-l63-f1	19	265 051	74 097	3,577081
Sum-l47-f1	10	8619	3505	2,459058
Sum-l90-f1	10	12 650	5154	2,454404
Sum-l45-f1	10	30 418	15 204	2,000658

4. Discussion

This is the first model comparison study of automatic versus manual anatomical brain image segmentation that comprehensively evaluates and classifies discrepancies on a per-label basis. The strength of this method is that it could be applied for other atlas databases. In this project the software application ITK-SNAP and complementary application c3d were used for image processing and the software MIRTK was used for image registration. MIRTK offers similar image processing tools as ITK-SNAP and c3d and could have been used for the image processing as well. The reason MRITK was not used for the whole process was because the idea of image registration arose when the idea of creating probability maps emerged which was after the data processing had been completed.

The limitation of this method was the image processing. Since the model comparison was on a per-label basis, a lot of data was created during the image processing which was very time-consuming. Another limitation was the command to split image components into binary images of each component. The command for this had an upper limit of 254, meaning that an image could only be divided into a maximum of 254 binary images. This forced other solutions to determine the total number of discrepancies that occur in a label to be found. Even though not all individual discrepancies were analysed in more detail, information still got lost due to this inconvenience. For further projects, this may be an even bigger issue and should be taken into consideration.

4.1 Surface discrepancies

Discrepancies with both error type were found in all regions and atlases. Most of the discrepancies were due to the manual segmentation and were found in region 17 and 18. One reason for the fact that more discrepancies occur for manual segmentation then for the automatic segmentation, is that the expert can consider more detailed information about the anatomy than the automatic algorithm. The anatomy is not identical for all individuals and for some there may be larger deviations. The automatic approach is more general and can only consider anatomical characteristics that represented in the atlas. For example, this could be the case for the discrepancies presented in figure 5A and 5B. These discrepancies were very large and widely extended. If this is assumed to be a mistake, it would mean that the expert had segmented a large region incorrectly, which is unlikely. The decision behind the segmentation may be due to an anatomical deviation that the automatic algorithm could not consider.

The following characteristics were determined for each individual discrepancy component image: atlas number, label number, error type, voxel count, centroid coordinates, and extents in the x, y, and z directions. Data about the extents' values were collected with the idea that this could be used to investigate if there were some regions that deviated from the average border for this region. This was not further investigated because it was considered to be outside the scope of this project, but is something that may be interesting to study further.

The visual comparison was conducted for the five largest discrepancies in label 22 and 23 in atlas 1, 5, and 30. These labels were chosen because discrepancies with some of the largest volumes occurred in these regions. The discrepancies were due to both manual and automatic segmentation. During the visualization, the discrepancies for each region were not analysed for every atlas, because that work would be very time intensive and did not fit the timeframe for this project. The selection of atlases was also based on the location of large discrepancies. Overall, the shape of the analysed discrepancies could be described by the structure definitions in the error types GSL and LPF but apart from determining the structure, no other conclusion could be drawn due to the limited time frame of this project. The already existing typology does subscribe a wide range of shapes but it was created using the manual segmentations as a reference, leading to the next reason why the discrepancies were visually examined. Apart from qualitatively characterizing the discrepancies, they were visually examined to see if a new typology could be created or one that could be complementary to the already existing one and applied on investigations like this. But the visual analysis was very subjective, and a lot of personal interpretation was involved. More time would be needed for this in order to visually analyse more discrepancies to get better results and to achieve the second purpose I was aiming for.

4.2 Label holes

Label holes were found in both manual and automatic segmentations. Holes occurred more frequently in the automatic segmentations, but the holes found in the manual segmentations were of larger volumes. These results seem reasonable since it was assumed that holes are flaws and it is less likely that the expert has made such mistakes than the automatic algorithm. Furthermore, it is expected that the holes in the manual segmentation were much larger than the holes located in the automatic segmentations because if voxels were labelled as another voxel value than the surrounding voxels (creating a hole) in the manual segmentation, the expert must have had a reason for this and thought that these voxels did not belong to the surrounding regions. For the automatic algorithm not all disparities can be considered. The most holes were found in label 23 (right occipital lobe), both for the manual and automatic segmentations. This region is quite large and located posterior to the right parietal and temporal lobe. A reason so many holes have been segmented in this region may be because of its location, that many sulci run deep in this region but also close to a ventricle. This may increase the possibility that the segmentation methods have determined voxels as the background i.e. non-brain portions of the image or brain regions not included in the protocol. Another reason for the high number of holes can be the region's large size. It seems that a larger label volume has a higher possibility of a discrepancy occurring. Due to the timeframe of the project, further investigation of this result has not been conducted.

The goal of the visual analysis of the label holes was to determine if some in fact were not flaws. The basic beforehand assumption was that holes should be regarded as a flaw and would be a few voxels big. It was not expected to find such large holes and when it was found that the largest holes occurred due to manual segmentation, the urge to do a visual analysis arose. The 10 largest holes due to manual and automatic segmentation were chosen for visual analysis. Because it seemed as if the expert had good reasons for segmenting the holes where they did, 5 additional holes due to manual segmentation but with much smaller volume were analysed to investigate if there were more holes that had been incorrectly labelled. It is more logical that holes with bigger volume are less likely to occur as a flaw than smaller holes, because it is less likely that the expert have done such a big mistake. If a large hole occurs, it also means that the structure the expert thought would not belong to the region in question was also large. A larger structure is more visible and easier to interpret than small structures. No additional holes due to automatic segmentation were further analysed as the largest holes for this error type were of 7 voxels.

Overall, the holes due to manual segmentation were easier to determine than the holes due to automatic segmentation, much due to the advantages that manual holes had substantially larger volumes and were easier to interpret. A pattern was discovered during the visual analysis. When a hole was correctly segmented, it had been assigned a voxel value 0 and in cases where the hole was incorrectly segmented, the voxels forming the hole were assigned the value corresponding to the neighbouring region. It is reasonable that the true holes were labelled with 0 because they were located in a region part of a sulcus or ventricle with CSF. Furthermore, when analysing holes with smaller volume it was noticed that some holes occur within a region and some close to the border of neighbouring region which the hole has been assigned to. Regardless, a hole within a region not part of an area including CSF were assumed to be wrong. But in these cases, I started to reflect on whether it is the voxels forming the hole that have been wrongly labelled or if it is the few voxels separating the hole from the neighbouring region that has been wrongly labelled. This was considered even more when a hole (a14-h-aut-l30-c1) was found near its assigned neighbouring region, surrounded by other smaller holes belonging to that neighbouring region (see Figure 14).

4.3 Probability maps

Probability maps of the discrepancies were created to collect information about the frequency relationship between the discrepancy voxels across the 30 atlases and give an overview of which discrepant voxels that occur most often. The idea was that this could be used as a basis for determining whether certain discrepant voxels have been segmented correctly or not. If a discrepant voxel has a high voxel value in the probability map, one could argue that this voxel has been correctly segmented.

5. Conclusion

Automatic and manual brain images segmentation have been compared to investigate discrepancies between the segmentation methods to try to decide whether they can be attributed to flaws in the automatic segmentation or in the manual segmentation, and furthermore conclude if there are general rules that enable these decisions. This project has resulted in a model that enable extraction of the discrepancies between the manual and automatic segmentation into individual components for a quantitative characterization on a per-label basis. Discrepancies both on the surface and within a label, forming holes were found and analysed. Holes that were true were assigned voxel value 0 and labelled as part of a sulcus and ventricle while holes that have been incorrectly segmented were assigned with a voxel value of a neighbouring region. Furthermore, a visual analysis was carried out with the goal to characterize the shape of the discrepancies, but further investigations are needed to get better results. Probability maps of the discrepancies have been created and can be used as a basis for determining the probability that a certain discrepant voxel have been segmented correctly or not.

6. Acknowledgement

I would like to thank my supervisor Rolf A. Heckemann for the guidance during the project but also for being tremendously supportive and patient. I would also like to thank my family and friends for all the positive encouragement.

Reference list

1. Heckemann RA, Hajnal JV, Aljabar P, Rueckert D, Hammers A. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. Neuroimage. 2006;33(1):115-26.
2. Heckemann RA, Keihaninejad S, Aljabar P, Rueckert D, Hajnal JV, Hammers A. Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation. Neuroimage. 2010;51(1):221-7.
3. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Medical Imaging. 2015;15(1):29.
4. Zhang H, Fritts JE, Goldman SA. Image segmentation evaluation: A survey of unsupervised methods. Computer Vision and Image Understanding. 2008;110(2):260-80.
5. Toga AW, Thompson PM. The role of image registration in brain mapping. Image Vis Comput. 2001;19(1-2):3-24.
6. Zitová B, Flusser J. Image registration methods: a survey. Image Vis Comput. 2003;21(11):977-1000.
7. Oliveira FPM, Tavares JMRS. Medical image registration: a review. Computer Methods in Biomechanics and Biomedical Engineering. 2014;17(2):73-93.
8. Rueckert D, Sonoda LI, Hayes C, Hill DLG, Leach MO, Hawkes DJ. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Transactions on Medical Imaging. 1999;18(8):712-21.
9. Hammers Atlas Database [Internet]. Hammers A, Allom R et al., Hum Braim Mapp 2003 for regions 01-49, Gousias IS et al, Neuroimage 2008 for regions 50-83, Faillenot I et al. 2017 for regions 86-95. Available from: www.brain-development.org
10. Hammers A, Allom R, Koepp MJ, Free SL, Myers R, Lemieux L, et al. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum Brain Mapp. 2003;19(4):224-47.

11. Gousias IS, Rueckert D, Heckemann RA, Dyet LE, Boardman JP, Edwards AD, et al. Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. Neuroimage. 2008;40(2):672-84.
12. Wild HM HR, Studholme C, Hammers A Gyri of the human parietal lobe: Volumes, spatial extents, automatic labelling, and propabilistic atlases. PLoS ONE 12(8): e0180866. 2017.
13. Faillenot I, Heckemann RA, Frot M, Hammers A. Macroanatomy and 3D probabilistic atlas of the human insula. Neuroimage. 2017;150:88-98.
14. Yaakub SN, Heckemann RA, Keller SS, McGinnity CJ, Weber B, Hammers A. On brain atlas choice and automatic segmentation methods: a comparison of MAPER & FreeSurfer using three atlas databases. Scientific Reports. 2020;10(1):2837.
15. Linear ICBM Average Brain (ICBM152) Stereotaxic Registration Model [Internet]. Copyright (C) 1993-2009 Louis Collins, McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University. Available from: http://nist.mni.mcgill.ca/?p=798.
16. Mazziotta J, Toga A, Evans A, Fox P, Lancaster J, Zilles K, et al. A four-dimensional probabilistic atlas of the human brain. J Am Med Inform Assoc. 2001;8(5):401-30.
17. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage. 2006;31(3):1116-28.
18. ITK-SNAP. Convert3D Documentation ITK-SNAP [updated 2015-04-29 cited 2020 5/12]. Available from: http://www.itksnap.org/pmwiki/pmwiki.php?n=Convert3D.Convert3D.

Appendix

Appendix 1

Code to extract surface discrepancies from the terminal

```
#split the holefilled segmentations into binary images for each region, manual and
automatic separately
for type in man aut ; do
    c3d $bd/$type-hf.nii.gz -split -oo $wd/$type-l%01d.nii.gz
done

#the adding and separation process
for l in {1..95} ; do
    c3d $wd/man-l$l.nii.gz $wd/aut-l$l.nii.gz -scale 2 -add -replace 3 0 -split -oo
$wd/dc-l$l-f%0d.nii.gz
    for f in 1 2 ; do   #the extracting discrepancy process
        c3d $wd/dc-l$l-f$f.nii.gz -comp -dup -threshold 1 254 1 0 -multiply -split -oo
$wd/dc-l$l-f$f-c%00d.nii.gz  #duplicate the images, threshold all voxel value between
1 to 254 in one of the images, multiplying the images together before splitting into
discrepancies
        rm $wd/dc-l$l-f$f-c0.nii.gz  #remove all *-c0-files, that only consist of the
background
    done
done
```

Appendix 2

A list of the discrepancies chosen for the visual analysis of surface discrepancies. Table columns provide the file name, centroids coordinates (c_i, c_j, c_k (voxel), c_x, c_y, c_z (mm)), volume and extent voxels values. The file name describes in which atlas and region the discrepancy is found together with the error type and if it is the largest, second et cetera connected compound.

Name	c_i	c_j	c_k	c_x	c_y	c_z	Vol	Ext_i	ext_j	ext_k
a1-dc-l22-f1-c1	121,75	58,4217	73,6327	-31,7496	-67,5783	1,63273	26618	59	85	78
a1-dc-l22-f1-c2	131,476	24,0366	76,4146	-41,4756	-101,963	4,41463	82	10	9	11
a1-dc-l22-f1-c3	124,955	22,1194	81	-34,9552	-103,881	9	67	8	9	4
a1-dc-l22-f1-c4	115,909	27,4242	106,318	-25,9091	-98,5758	34,3182	66	7	11	8
a1-dc-l22-f1-c5	134,831	34,3898	105,932	-44,8305	-91,6102	33,9322	59	12	9	5
a1-dc-22-f2-c1	122,756	40,445	115,879	-32,7559	-85,555	43,8789	2782	32	28	18
a1-dc-22-f2-c2	110,955	51,033	81,3031	-20,9546	-74,967	9,30309	485	16	15	27
a1-dc-22-f2-c3	102,213	34,5603	113,348	-12,2128	-91,4397	41,3475	282	8	11	16
a1-dc-22-f2-c4	138,078	57,2941	61,4667	-48,0784	-68,7059	-10,5333	255	15	8	9
a1-dc-22-f2-c5	102,28	18,7702	71,9565	-12,2795	-107,23	-0,0434783	161	9	12	11
a1-dc-l23-f1-c1	51,5533	71,5967	57,4985	38,4467	-54,4033	-14,5015	13321	47	67	44
a1-dc-l23-f1-c2	75,7208	31,5324	87,2225	14,2792	-94,4676	15,2225	3195	21	52	52
a1-dc-l23-f1-c3	49,3813	44,8012	119,039	40,6187	-81,1988	47,0385	493	18	16	12
a1-dc-l23-f1-c4	62,1178	42,0356	123,447	27,8822	-83,9644	51,4467	450	21	17	17
a1-dc-l23-f1-c5	30,2462	56,2462	84,3205	59,7538	-69,7538	12,3205	390	16	21	15
a1-dc-23-f2-c1	57,5669	56,5709	100,475	32,4331	-69,4291	28,4753	2981	38	31	48
a1-dc-23-f2-c2	67,5761	33,6587	58,0979	22,4239	-92,3413	-13,9021	2892	27	37	20
a1-dc-23-f2-c3	75,7652	34,4534	116,032	14,2348	-91,5466	44,0324	247	7	10	20
a1-dc-23-f2-c4	67,2254	28,5915	109,62	22,7746	-97,4085	37,6197	71	12	4	9
a1-dc-23-f2-c5	55,403	64,5821	67,6716	34,597	-61,4179	-4,32836	67	10	10	4
a5-dc-l22-f1-c1	128,414	55,5131	97,2445	-38,4138	-70,4869	25,2445	26981	67	59	78
a5-dc-l22-f1-c2	129,404	36,5426	102,011	-39,4043	-89,4574	30,0106	94	13	7	14
a5-dc-l22-f1-c3	138,679	41,0494	60,5185	-48,679	-84,9506	-11,4815	81	8	16	8
a5-dc-l22-f1-c4	138,525	40,3729	92,5593	-48,5254	-85,6271	20,5593	59	11	5	14
a5-dc-l22-f1-c5	107,327	51,6538	82,6923	-17,3269	-74,3462	10,6923	52	8	5	15
a5-dc-l22-f2-c1	57,5669	56,5709	100,475	32,4331	-69,4291	28,4753	2981	38	31	48
a5-dc-l22-f2-c2	67,5761	33,6587	58,0979	22,4239	-92,3413	-13,9021	2892	27	37	20
a5-dc-l22-f2-c3	75,7652	34,4534	116,032	14,2348	-91,5466	44,0324	247	7	10	20
a5-dc-l22-f2-c4	67,2254	28,5915	109,62	22,7746	-97,4085	37,6197	71	12	4	9
a5-dc-l22-f2-c5	55,403	64,5821	67,6716	34,597	-61,4179	-4,32836	67	10	10	4
a30-dc-l23-f1-c1	60,799	69,8748	58,9355	29,201	-56,1252	-13,0645	2124	33	42	18
a30-dc-l23-f1-c2	76,382	33,8714	97,4984	13,618	-92,1286	25,4984	1563	28	22	47
a30-dc-l23-f1-c3	76,7655	20,1219	78,1348	13,2345	-105,878	6,13482	853	23	11	38
a30-dc-l23-f1-c4	71,3169	50,5432	81,3498	18,6831	-75,4568	9,34979	243	13	10	28
a30-dc-l23-f1-c5	50,5697	51,2061	51,3455	39,4303	-74,7939	-20,6545	165	17	14	3
a30-dc-l23-f2-c1	53,8407	51,2035	92,6522	36,1593	-74,7965	20,6522	24134	47	34	73
a30-dc-l23-f2-c2	71,3649	31,7014	63,5346	18,6351	-94,2986	-8,4654	1055	16	19	24
a30-dc-l23-f2-c3	63,4186	51,3256	53	26,5814	-74,6744	-19	43	10	10	1
a30-dc-l23-f2-c4	66,2353	60,0588	55,8529	23,7647	-65,9412	-16,1471	34	8	5	3
a30-dc-l23-f2-c5	64,0625	56,1875	54	25,9375	-69,8125	-18	16	9	4	1

Appendix 3

A list of the discrepancies chosen for visual analysis of label holes. Table columns provide the file name, centroids coordinates (c_i, c_j, c_k (voxel), c_x, c_y, c_z (mm)), volume and extent voxels values. The file name describes in which atlas and region the discrepancy is found, together with the error type and if it is the largest, second et cetera connected compound.

Name	c_i	c_j	c_k	c_x	c_y	c_z	Vol	Ext_i	Ext_j	Ext_k
a6-h-man-l18-c1	97,0571	48,6857	46,3143	-7,05714	-77,3143	-25,6857	70	3	8	6
a2-h-man-l58-c1	100,915	137,366	117,62	-10,9155	11,3662	45,6197	71	2	23	5
a2-h-man-l14-c1	157,28	104,82	56,72	-67,28	-21,18	-15,28	50	8	5	4
a25-h-man-l33-c1	50,1856	66,1959	90,7216	39,8144	-59,8041	18,7216	97	13	4	8
a25-h-man-l23-c1	66,0082	53,418	78,8197	23,9918	-72,582	6,81967	122	5	8	8
a24-h-man-l22-c1	110,84	48,0922	77,8447	-20,8398	-77,9078	5,84466	206	11	13	8
a1-h-man-l85-c1	28,9178	95,4658	107,562	61,0822	-30,5342	35,5616	73	8	9	7
a16-h-man-l22-c1	116,346	38,7115	89,8462	-26,3462	-87,2885	17,8462	52	4	8	8
a11-h-man-l22-c1	118,083	47,8333	96,6944	-28,0833	-78,1667	24,6944	72	8	8	5
a10-h-man-l23-c1	60,3924	49,5316	56,4557	29,6076	-76,4684	-15,5443	79	13	2	4
a25-h-aut-l84-c1	131,429	87,7143	114	-41,4286	-38,2857	42	7	2	2	3
a24-h-aut-l61-c1	82,7143	87,2857	143,286	7,28571	-38,7143	71,2857	7	3	2	2
a10-h-aut-l64-c1	113,833	62,1667	60	-23,8333	-63,8333	-12	6	3	2	3
a16-h-aut-l57-c1	40,6	142	95,8	49,4	16	23,8	5	2	1	3
a8-h-aut-l57-c1	35	144	74,25	55	18	2,25	4	3	1	2
a5-h-aut-l23-c1	55,5	60,5	105	34,5	-65,5	33	4	2	2	1
a26-h-aut-l50-c1	144	131,5	103,5	-54	5,5	31,5	4	1	2	2
a1-h-aut-l29-c1	43,75	131,5	120,25	46,25	5,5	48,25	4	2	2	2
a17-h-aut-l51-c1	41,5	128	113,5	48,5	2	41,5	4	2	1	2
a13-h-aut-l30-c1	122	64,25	56	-32	-61,75	-16	4	3	2	1
a28-h-man-l51-c1	63,5455	119,545	115,909	26,4545	-6,45455	43,9091	11	7	2	2
a28-h-man-l50-c1	96	99,7273	138,364	-6	-26,2727	66,3636	11	1	3	6
a19-h-man-l22-c1	125,091	52,2727	59	-35,0909	-73,7273	-13	11	5	3	1
a15-h-man-l17-c1	86,4545	60,3636	55,3636	3,54545	-65,6364	-16,6364	11	2	2	4
a14-h-man-l22-c1	118,636	39,4545	92,8182	-28,6364	-86,5455	20,8182	11	3	2	5