 doi:10.20944/preprints202204.0068.v1

Article

Functional Data Analysis for imaging mean function estimation: computing times and parameter selection

Juan Arias-López [1,2]*, Carmen Cadarso-Suárez[1,2] and Pablo Aguiar-Fernández [3,4]

[1] Biostatistics and Biomedical Data Science Unit. Department of Statistics, Mathematical Analysis, and Operational Research, Universidade de Santiago de Compostela, Spain
[2] CITMAga, 15782 Santiago de Compostela, Spain
[3] Nuclear Medicine Department and Molecular Imaging Group, University Clinical Hospital (CHUS) and Health Research Institute of Santiago de Compostela (IDIS), Santiago de Compostela, Spain
[4] Molecular Imaging Group, Department of Psychiatry, Radiology and Public Health, Faculty of Medicine, Universidade de Santiago de Compostela, Spain
* Correspondence: juanantonio.arias.lopez@usc.es

Abstract: Functional Data Analysis (FDA) is a relatively new field of statistics dealing with data expressed in the form of functions. FDA methodologies can be easily extended to the study of imaging data, an application proposed in previous publications where the authors settle the mathematical groundwork and properties of the proposed estimators. This methodology allows for the estimation of mean functions and simultaneous confidence corridors (SCC), also known as simultaneous confidence bands, for imaging data and for the difference between two groups of images. This is especially relevant for the field of medical imaging, as one of the most extended research setups consists on the comparison between two groups of images, a pathological set against a control set. FDA applied to medical imaging presents at least two advantages compared to previous methodologies: it avoids loss of information in complex data structures and avoids the multiple comparison problem arising from traditional pixel-to-pixel comparisons. Nonetheless, computing times for this technique have only been explored in reduced and simulated setups. In the present article, we apply this procedure to a practical case with data extracted from open neuroimaging databases and then measure computing times for the construction of Delaunay triangulations, and for the computation of mean function and SCC for one-group and two-group approaches. The results suggest that previous researcher has been too conservative in its parameter selection and that computing times for this methodology are reasonable, confirming that this method should be further studied and applied to the field of medical imaging.

Keywords: Functional Data Analysis; Image Processing; Brain Imaging; Neuroimaging; Computational Neuroscience; Data Science

1. Introduction

1.1. Functional Data Analysis

The field of statistics involved in the mathematical development of tools for the analysis of data in the form of functions is known as Functional Data Analysis (FDA). From a FDA scope, the minimum unit of data to be analyzed is not one data point itself, but rather a function which, usually, corresponds to a single participant in a biomedical study or, in more complex scenarios, a series of functions assigned to each of the participants.

The area of FDA is still underdeveloped and much research with new applications appears every year in scientific journals. However, although a strict definition of the field is not established - nor it appears as desirable - there are a series of characteristics which appear to be inherent to functional data and which can be helpful to understand the methods and objectives within the scope of this field. First, functional data are continuously defined and, as such, single instances of functional data are considered mostly irrelevant and just as realizations of the underlying function with is the main object of analysis. This is a necessary constraint established in order to work with this data using the computational tools available to us. Second, the basic element of the analytical process performed in FDA

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2022 doi:10.20944/preprints202204.0068.v1

is the whole function itself, and not the individual data elements of which it is composed. Finally, functional data usually appears associated to some sort of temporal variable and it is also assumed to have some regularity conditions [1].

Taken together, functional data usually consists of a sample of independent functions with values which are located in a compact and predefined grid or interval (I) and are, in most of the cases, assumed to exist in a Hilbert space (L^2):

$$X_1(t), X_2(t), ..., X_n(t); I = [0, T] \in L^2 \tag{1}$$

In the last years, FDA has gained momentum evidenced by a rise in popularity in several applied research areas and the publication of multiple works including monographs [1] and review articles [2]. Now that this knowledge is available to the public and FDA's theoretical basis and applications are beginning to be established, researchers are starting to consider the use of FDA tools for extended setups such as its application in the field of medical imaging.

1.2. Applicability of FDA to Imaging Data

In the context of biomedicine, there is great interest in medical imaging data such as the ones obtained from brain scanners, images of tumor tissues, among others [1]. Nevertheless, smoothing methods proposed in the scientific literature to date which are focused on imaging data(e.g. kernel smoothing, tensor product smoothing...) suffer from a severe problem of *leakage* for high-complexity data structures, showing difficulties carrying out estimations in boundary regions and thus resulting in inappropriate smoothing.

In addition, there are other problems aside from estimations of the value for a single point when analysing medical imaging data with traditional methods. Another problematic arises for the estimation of the associated uncertainty of that estimation (i.e. its confidence band), a problem which becomes even more complicated when considered that also the spatial correlation has to be taken into account. So far, the predominant techniques for mean imaging data estimation and also for the computation of associated uncertainty have been the methodologies termed as *mass univariate approaches*. From this mass univariate approach, every pixel in an image is considered as independent, then a pixel-to-pixel comparison is performed with classical methods such as *T-tests*. The associated multiple comparison problem is then solved applying popular approaches such as the Bonferroni correction or the application of random field theory [3], which are *ad hoc* corrections very dependent on the chosen threshold.

These problems associated with classical methods for mean estimation are avoided by the FDA technique proposed by Wang et al. [4]. In this article, the authors propose a way to avoid the problem of *leakage* on complex data structures using bivariate splines over Delaunay triangulations (see Section 2.2), thus preserving the most complex and important details of imaging data structures. Besides, the proposed methodology considers imaging data as an instance of functional data which is continuously defined (as explained in Section 1.1) but observed on a regularly defined grid. Given that the imaging data is treated as functional data, attention naturally moves from the pixel as the minimal analytic unit to the analysis of images as a whole. This allows not only for the calculation of the mean function of a group of images, but also for the estimation of simultaneous confidence corridors (SCC; also known as *simultaneous confidence bands*), an approach which has been proven superior to conventional multiple comparison approaches [5]. Further, Wang and collaborators [4] also describe the proposed bivariate spline estimators, test their asymptotic properties, describe the attributes of SCC based on these estimators, and extract coverage probability for the obtained mean function. The conclusion of the article is that the proposed SCC methodology accounts for the correct probability coverage both in one-group and two-group comparison setups.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2022 doi:10.20944/preprints202204.0068.v1

However, although the proposed methodology accounts for the correct probability coverage, the computational resources necessary for its application are not addressed by the authors and thus its utility for a practical case is yet to be fully understood. Previous research [6] has tested this methodology in limited setups, concluding not only that the parameters proposed by the authors were too conservative, but also that the amount of time to obtain results was in the tolerable range for modern computational capabilities. This suggests that moving towards a FDA setup in studies comparing groups of medical images might be a sensible thing to do, however, this study tested the herein studied FDA methodology with simulated data which was not very complex in its structure and was also estimated with predefined parameters. For these reasons, there is a necessity for testing computing times of this method in a practical case, with real imaging data and a higher number of patients.

1.3. Objectives

Given that the computational costs for this methodology are not fully explored and that the only available results - although promising - were only applied to simulated data [6], we consider that this article's objective consists on testing the practical utility of this novel methodology by studying the computational efforts necessary to implement it for a practical case with data obtained from open brain imaging databases. We perform this analysis by evaluating computing times for the calculation of the polygonal domain of the Delaunay triangulations necessary to carry out this method and also evaluating computing times for the calculation of mean functions of a group of images (one-group setup) and for the comparison between two groups in order to highlight areas with differences in brain activity (two-group setup).

2. Materials and Methods

2.1. Imaging Data

There are different approaches to brain imaging, resulting on data with differential peculiarities. For this case, we chose to use 18F-FDG Positron Emission Tomography (PET) data given its reliability [7] and extended use in clinical neuroscience. In this imaging technique, Fluorodeoxyglucose (18F-FDG), a radioisotope analogue of glucose, is used as tracer to monitor brain metabolic rates. Positron emission rates by molecules of 18F-FDG trapped in brain tissues are used as an indirect measure of glucose consumption which is then reconstructed producing 3D images for the position of this tracer in the brain.

We drawn upon the Alzheimer's Disease Neuroimaging Initiative [8], selecting 18F-FDG PET data for a control group (75 patients; 44 male; age: 75.56 ± 4.96 years) and a Alzheimer's Disease (AD) group (51 patients; 30 male; age: 74.03 ± 7.25 years) summing 126 participants. A critical step in any neuroimaging study is the existence of a precise point-to-point correspondence when comparing scans from brains which present unique shape and size. For this reason, images were realigned, unwrapped, co-registered with MRI data, spatially normalized, mean proportionally scaled, and masked following standard procedures deployed by Statistical Parametric Mapping (SPM) software [9]. As a result, the data used for this study is treated following standard procedures for brain imaging research and undergoes a pre-processing workflow, guaranteeing pixel-to-pixel correspondence before the application of our examined technique.

2.2. Delaunay Triangulations

Delaunay triangulations consist on multiple triangles created by the union of vertices in which no vertices falls inside the circumcircle of a given triangle. The FDA approach we are examining uses bivariate splines over a pre-existing grid of these triangulations specifically designed for the shape of the objective image to analyse. This approach reduces the loss of information in boundary regions and thus enhances the obtained results' accuracy. In order to calculate these triangulations, we use the *Triangulation* R package [10] on a slice of brain imaging data and test its computing costs for growing values of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2022 doi:10.20944/preprints202204.0068.v1

triangulation fineness degree. An example of these triangulations for our practical case can be seen in Figure 1 and computing times are analysed in Section 3.1.

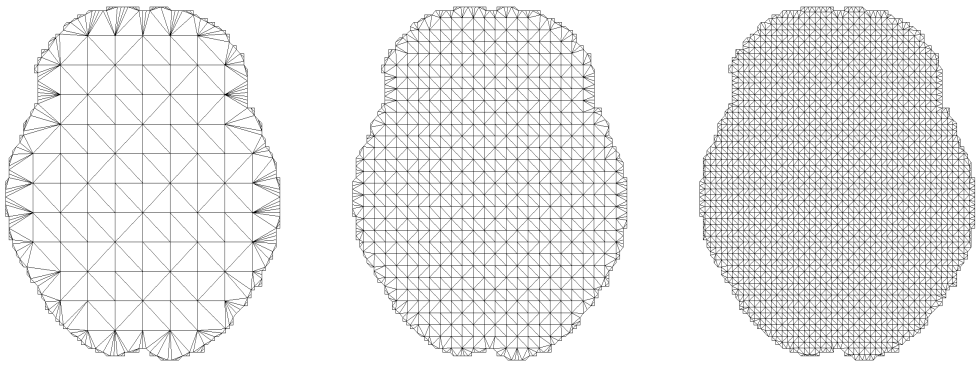

Figure 1. Delaunay triangulations produced for our practical case with real brain imaging data. Increasing triangulation's degree of fineness is measured by parameter N. (**a**) N=10. (**b**) N=25. (**c**) N=50.

2.3. Mean Function and SCC for one-group setup

The proposed FDA methodology allows for two different calculations: the estimation of a group of images' mean function and its associated SCC in the form of images, and the comparison between two groups of images in order to obtain the mean function for the difference between groups. In this subsection we compute one-group mean function for a group of images together with its associated SCC for a given *alpha* value with the help of functions implemented in *ImageSCC* R package [11]. Examples of the results obtained using this methodology can be found in Figure 2. Computing times for different triangulation fineness degrees are analysed in Section 3.2.

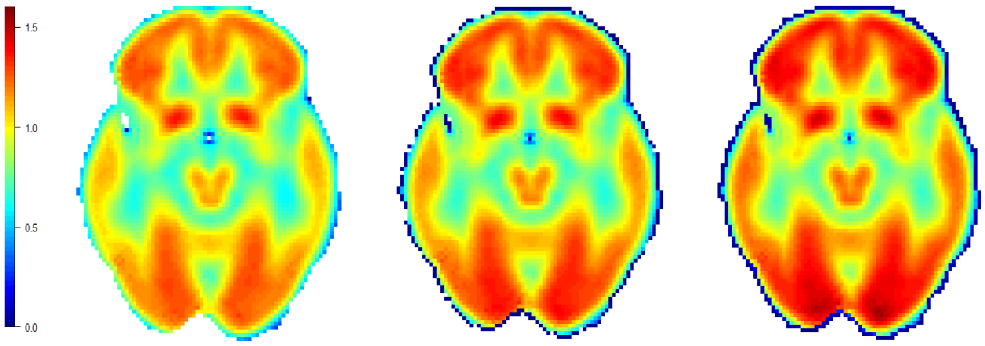

Figure 2. Example of (**a**) Lower SCC; (**b**) Mean Function; and (**c**) Upper SCC for brain imaging data. SCCs calculated for $\alpha = 0.05$ using Delaunay triangulations (fineness degree $N = 10$).

2.4. Mean Function and SCC for two-group setup

This technique can be extended to a two-sample setup in which the mean function for the difference between groups of images is obtained, together with their SCC. Using this information, we can also calculate which regions of the image present activity patterns

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2022 doi:10.20944/preprints202204.0068.v1

falling outside expected values, suggesting a significant difference in activity for that region in one group compared to another. Results can be found in Figure 3 and computing times are analysed in Section 3.3.

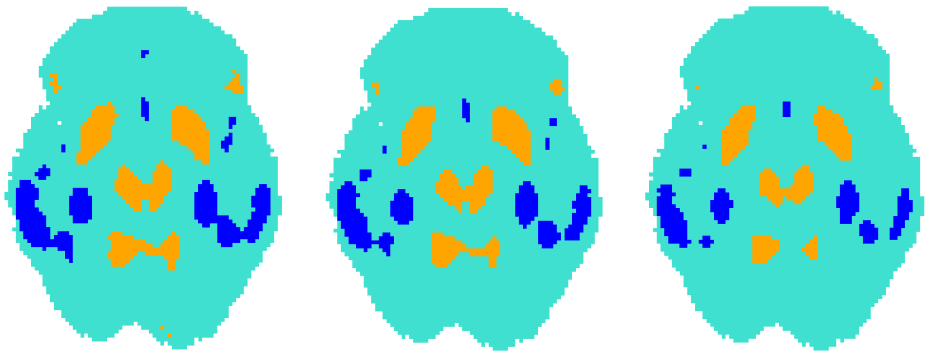

Figure 3. Example of results for a two-sample approach. Blue indicates detected hypo-activity while orange indicates hyper-activity. Delaunay triangulations' fineness degree $N = 10$. (**a**) $\alpha = 0.1$. (**b**) $\alpha = 0.05$. (**c**) $\alpha = 0.01$.

3. Results

In this section we proceed to summarize the obtained results for the methodologies described in Section 2.2, Section 2.3, and Section 2.4 when applied to real neuroimaging data (see Section 2.1) using Delaunay triangulations with a growing degree of fineness, which is the main tuning parameter for these approaches as they indicate the degree of complexity for the grid upon which the data is analysed.

3.1. Delaunay Triangulations

In Figure 4 we examine computing times for the generation of a grid of Delaunay triangulations for highly-complex data structures such as the ones used in this case, these triangulations are the basis for imaging mean function and SCC estimation and will be used in the following sections to evaluate computing times for the two possible applications of this methodology. Our results, together with previous findings using simulated data [6] suggest that the degree of triangulation fineness proposed in the scientific literature (*N=8*) [4] is too conservative and that modern computers available to data science researchers can easily handle the computation of triangulation grids for N values of up to *N=25* and higher. However, as described in the following sections, the growing complexity of these grids causes an accumulative effect reflected in the computing times of mean functions for groups of images, which has to be taken into account.

3.2. One-group Mean Function and SCC Estimation

In Figure 5 we provide a graphical summary of computing times to obtain one-group mean function and associated SCC in the form of images (as shown in Figure 2). Aside from the triangulation degree of fineness, which grows in order to test to what extent the polygonal domain affects the costs of computation, estimated mean function and SCC were calculated using parameters recommended by Wang et al. [4] including: degree of bivariate spline for mean estimation $d.est = 5$, degree of bivariate spline for construction of SCC $d.band = 2$, smoothness parameter $r = 1$, and a vector of candidates for penalty parameter with values ranging from 10^{-6} to 10^{3}.

We can see that computing times for this process remain fairly stable in the range between one to five hours of processing for triangulations with a fineness degree below $N = 15$. Above that value, computing times start to rise linearly, resulting in 22 hours of processing for triangulation's fineness degrees of $N = 25$. This is very relevant, as in previous sections (see Section 3.1) we considered $N = 25$ as sensible for the computation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2022 doi:10.20944/preprints202204.0068.v1

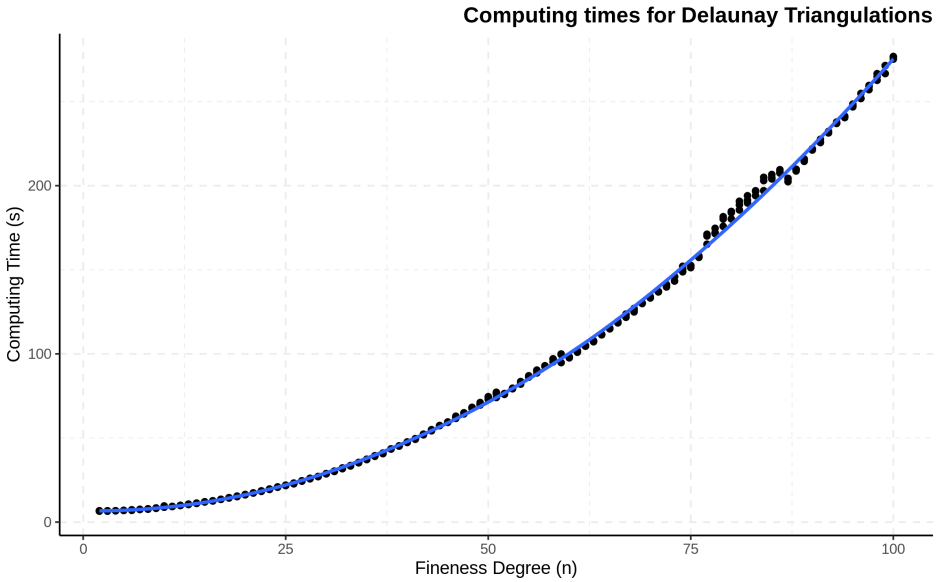

Figure 4. Computing times for Delaunay triangulations for complex neuroimaging data structures with growing fineness degree values. Curve fitted with local (LOESS) regression.

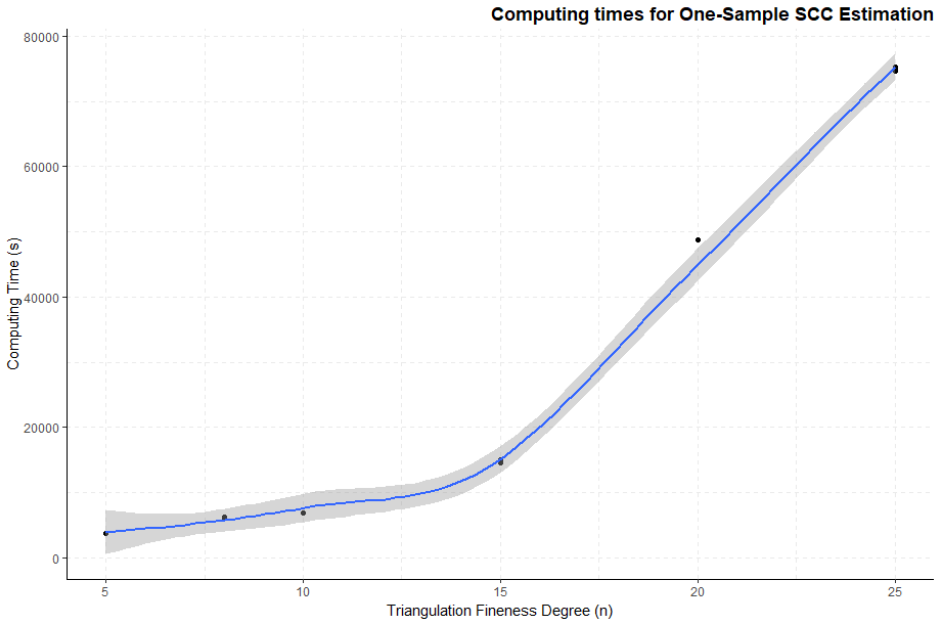

Figure 5. Computing times for one-group mean function and SCC estimation for neuroimaging data with growing value of triangulation fineness degree. Curve fitted using local (LOESS) regression.

of Delauney triangulation parameters, however, we can see that the cumulative effect of increasing the grid's complexity makes the application of this technique much more difficult and expensive in terms of time and computational power.

3.3. Two-group Mean Function and SCC Estimation

In Figure 6 we can visually examine computing costs for performing a two-group comparison using this FDA technique, which implies the calculation of mean function for the difference between both groups of images and also the associated SCC (as shown in Figure 3). We carry out this time calculation for a growing value of triangulation's fineness

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2022 doi:10.20944/preprints202204.0068.v1

degree using the same parameters recommended by Wang et al. [4] and described in the previous subsection.

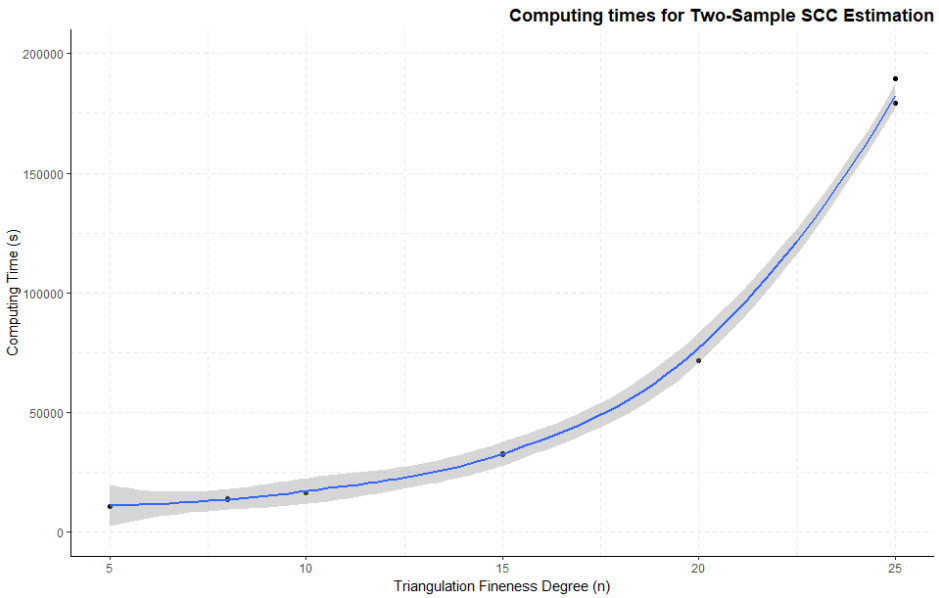

Figure 6. Computing times for two-group mean function and SCC estimation for the differences between groups with growing value of triangulation fineness degree. Curve fitted using local (LOESS) regression.

These results show a similar pattern to the ones presented in Figure 5. Computing times remain stable and sensible until a triangulation's fineness degree threshold placed approximately around $N = 15$. Above that value, computing times for this methodology grow and can even go above the 50 hours of time. It is important to note that computing times for the two-sample case, which is the most significant for clinical practice, are much higher than for the one-sample case. Besides, in line with the results of the previous subsection, it does not seem sensible to choose a triangulation fineness parameter only on the basis of triangulation's computing times. We need to take the whole process into consideration and that forces us to choose lower values (e.g. $N = 15$) as appropriate for computations inside a sensible time frame.

4. Discussion

The main goal for this article was to first implement Wang and colleagues' [4] FDA methodology for the estimation of mean functions and SCC for data in the form of images to a practical case using neuroimaging data. We approached this objective by gathering PET data from open neuroimaging databases focused, in this case, on AD and other dementias. After a complex pre-processing stage performed in order to guarantee pixel-to-pixel comparability, we proceeded to calculate the Delaunay triangulation polygonal space that serves as fundamental grid for estimations from a FDA approach. Likewise, we proceeded to estimate mean function and SCC for a single group of images (see Section 3.2) and for the difference between two groups of images (see Section 3.3), detecting this way areas with patterns of hypo- or hyper-activity in one group compared to another.

After carrying out these processes of triangulation (Figure 1), one-group mean and SCC estimation (Figure 2), and two-group mean difference and SCC estimation (Figure 3) there was still a necessity to evaluate whether the computational costs associated to this methodology are worthy of the results obtained. Although there are previous publications covering computing times and parameter selection for this methodology [6], these were performed with low-complexity simulated data. For this reason, in this article we aimed to perform the different stages of this FDA methodology using diverse triangulation parameters in order to assess computation times and also to provide future researchers

 doi:10.20944/preprints202204.0068.v1

with indications on the preference of choice when replicating or expanding these results with real applications.

The results obtained in this study suggest that, in line with previous publications [6] and against the default parameters suggested by Wang and colleagues [4], a sensible degree of fineness for the Delaunay triangle polygonal domain can be higher than $N = 8$. According to the visualizations presented for triangulation computing times in Figure 4, together with results displayed in Figures 5 and 6, this parameter of triangulation fineness can be increased to at least $N = 15$ and still obtain results inside sensible time limits when using computers with relatively high computation power (See Section 5). Besides, we can also see that computing times - both for one-sample and two-sample cases - grow as the triangulation grid's complexity increases, reaching critical points in which computing times start to be measured in days rather than hours. This goes in line with expected outcomes for functional data methodologies, which are meant to be applied to a high number of cases, whereas increases in the intricacy of the triangulation meshes tend to produce cumulative effects deriving in increased computing times due to the higher complexity of the calculations involved.

In summary, the proposal of applying FDA techniques to imaging data as bi-dimensional extensions of functional data is feasible and promising. We carried out the different steps necessary for a practical case application with brain imaging data, obtaining plausible results which go in line with previous literature in a sensible amount of time. We also suggest that, given the current computational power usually available at biomedical data science research groups, parameters for mean function and SCC estimation can be stricter than the ones suggested in previous articles. We also consider that appropriate choice of triangulation parameters is the most relevant decision for this methodology, as the cumulative effect of their complexity appear to be the most influential factor affecting computing times. In short, these results confirm the utility of FDA techniques for real practical cases of imaging analysis as they display desirable properties such as stability and reasonable computing times.

However, there is still a gap of knowledge to bridge with regards to this new methodology. Traditionally, SPM has been the golden standard for brain imaging studies. This software suite relies on simple statistical tests such as T-tests repeated following what is known as the *mass univariate approach*, then correcting false positives derived from multiple comparisons with methods such as Bonferroni's correction. Thus, SPM considers pixels as independent units inside the image, which are compared against its correspondent pixel in another set of images in order to conclude whether the value of brain activity in that coordinate is equal, higher, or lower than its counterpart. This approach can elude this problematic as FDA considers the whole image as the basic data unit and, besides, it can potentially obtain better results at complex data structures such as brain images. For these reasons, it is reasonable to argue that FDA should detect changes in brain activity more accurately than SPM and thus be more useful for clinical practice and research in fields such as neurodegenerative diseases' diagnosis and other fields of medical imaging which are of great relevance in this century. For this reasons, future research should strive to mathematically address the predictive value of this methodology compared to SPM, in order to have a clearer image of the advancement the implementation of this methodology could mean for researchers and clinical professionals in the field of neuroimaging and medical imaging more generally.

5. Computer Specifications

This study was carried out using the Biostatistics and Biomedical Data Science's server available at University of Santiago de Compostela, a computer with the following specifications and R version. Model: ProLiant DL160 Gen9; OS: Ubuntu 18.04.6 LTS x86; CPU: Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz; RAM memory: 118 Gb; R version: 4.0.3 (2020-10-10).

Supplementary Materials: Supporting information and scripts for replication of this study can be downloaded at our GitHub open repository.

Funding: This research received no external funding.

Data Availability Statement: 18F-FDG PET data for our practical case was drawn upon the Alzheimer's Disease Neuroimaging Initiative, a platform that collects data from different research institutions focusing on AD diagnosis.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

FDA	Functional Data Analysis
SCC	Simultaneous Confidence Corridor
PET	Positron Emissionn Tomography
18F-FDG	18-Fluorodeoxyglucose
AD	Alzheimer's Disease
SPM	Statistical Parametric Mapping

References

1. Ramsay, J. O. **(2004)**. Functional data analysis. *Encyclopedia of Statistical Sciences*, 4. http://dx.doi.org/10.1002/0471667196.ess3138
2. Wang, J.-L., Chiou, J.-M., & Müller, H.-G. **(2016)**. Functional data analysis. *Annual Review of Statistics and Its Application*, 3, 257–295. http://dx.doi.org/10.1146/annurev-statistics-041715-033624
3. Worsley, K.J., Taylor, J.E., Tomaiuolo, F., Lerch, J. **(2004)**. Unified univariate and multivariate random field theory. *NeuroImage* 23, S189–S195. http://dx.doi.org/10.1016/j.neuroimage.2004.07.026
4. Wang, Y., Wang, G., Wang, L., & Ogden, R. T. **(2020)**. Simultaneous confidence corridors for mean functions in functional data analysis of imaging data. *Biometrics*, 76(2), 427–437. http://dx.doi.org/10.1111/biom.13156
5. Degras, D.A. **(2011)**. Simultaneous confidence bands for nonparametric regression with functional data. *Statistica Sinica* pp. 1735–1765. http://dx.doi.org/10.5705/ss.2009.207
6. Arias-López, J. A., Cadarso-Suárez, C., & Aguiar-Fernández, P. (**2021**). Computational Issues in the Application of Functional Data Analysis to Imaging Data. *Computational Science and Its Applications – ICCSA 2021* (O. Gervasi, B. Murgante, S. Misra, C. Garau, I. Blečić, D. Taniar, B. O. Apduhan, A. M. A. C. Rocha, E. Tarantino, & C. M. Torre (eds.); pp. 630–638). Springer International Publishing. http://dx.doi.org/10.1007/978-3-030-86960-1_46
7. López-González, F. J., Silva-Rodríguez, J., Paredes-Pacheco, J., Niñerola-Baizán, A., Efthimiou, N., Martín-Martín, C., Moscoso, A., Ruibal, Á., Roé-Vellvé, N., & Aguiar, P. **(2020)**. Intensity normalization methods in brain FDG-PET quantification. NeuroImage, 222, 117229. http://dx.doi.org/10.1016/j.neuroimage.2020.117229
8. Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., ... & Beckett, L. **(2005)**. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimer's & Dementia, 1(1), 55-66.
9. Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., Nichols, T. E., Klebel, S. J., Nichols, T. E., Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., & Nichols, T. E. **(2006)**. Statistical Parametric Mapping: The Analysis of Functional Brain Images. Elsevier.
10. Lai, M.J., Wang, L.: Triangulation: Triangulation in 2D domain **(2020)**. R package version 0.1.0.
11. Wang, Y., Wang, G., Wang, L. **(2020).** ImageSCC: SCC for Mean Function of Imaging Data. R package version 0.1.0.