Stat

The ISI's Journal for the Rapid Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.39

Surface boxplots

Marc G. Genton[a]*, Christopher Johnson[b], Kristin Potter[d], Georgiy Stenchikov[a] and Ying Sun[c]

Received 03 December 2013; Accepted 17 December 2013

In this paper, we introduce a surface boxplot as a tool for visualization and exploratory analysis of samples of images. First, we use the notion of volume depth to order the images viewed as surfaces. In particular, we define the median image. We use an exact and fast algorithm for the ranking of the images. This allows us to detect potential outlying images that often contain interesting features not present in most of the images. Second, we build a graphical tool to visualize the surface boxplot and its various characteristics. A graph and histogram of the volume depth values allow us to identify images of interest. The code is available in the supporting information of this paper. We apply our surface boxplot to a sample of brain images and to a sample of climate model outputs. Copyright © 2014 John Wiley & Sons Ltd.

Keywords: band depth; fast algorithm; functional boxplot; image data; large dataset; ranking; visualization; volume depth

1 Introduction

As a result of new technologies and sophisticated monitoring devices, an increasing amount of functional data, including curves, surfaces, and images, is collected in many different fields of science and engineering, such as environmetrics, geophysics, biometrics, medicine, and neuroscience, to name a few. For example, modern brain imaging techniques, such as functional magnetic resonance imaging (fMRI), measure brain activities and produce brain images for the assessment of neurological disorders, and weather satellite images play an important role in weather forecasting. Besides observational functional data, experimental functional data generated by computer models, for example, various climate model outputs, have grown in size and complexity as well. Analysing and extracting useful information from such complex data have become challenging especially in higher dimensions. Therefore, functional techniques designed for surfaces or images are needed.

When sample surfaces or images are available, it is important to develop intuitive and efficient visualization tools to represent the data and highlight their characteristics to make the best use of the data resources. Computer-based visualization is widely used in many disciplines to help understand and communicate data, as well as to gain insights into the underlying processes. Many different methods and software have been developed for various purposes. For example, Walter et al. (2010) reviewed visualization methods for image data in biology, and medical image visualization was discussed by Blackwell et al. (2000) and McAuliffe et al. (2001).

[a]**CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia**
[b]**Scientific Computing and Imaging Institute, Salt Lake City, UT 84112, USA**
[c]**Department of Statistics, Ohio State University, Columbus, OH 43210, USA**
[d]**Department of Computer and Information Science, University of Oregon, Eugene, OR 97403, USA**
***Email: marc.genton@kaust.edu.sa**

(wileyonlinelibrary.com) DOI: 10.1002/sta4.39

The ISI's Journal for the Rapid Dissemination of Statistics Research

In this paper, from a statistical point of view, we aim to use descriptive statistics for sample surfaces or images to find the most representative sample surface or image as well as to detect potential outliers. This is statistically interesting but challenging, mainly because of three issues we are facing. First, for functional data analysis, as the entire surface or image is the information unit, we need a robust method to define the median surface or image and detect outliers. Second, a computationally efficient procedure is needed owing to the usually large volume of data for high-resolution surfaces or images. Third, an interactive and user-friendly visualization tool is desirable to display important statistics and data features.

To address these issues, as suggested by Sun & Genton (2011), we propose the surface boxplot in 3D based on the surface or image ranking induced by the notion of volume-based data depth. For computations, the fast algorithm developed by Sun et al. (2012) is adapted to compute the volume data depth values in 3D. Besides displaying important data features, we create an interactive visualization tool that allows users to understand the data better from different perspectives.

The remainder of our paper is organized as follows. The ranking of surfaces based on volume depth is described in Section 2, and the proposed surface boxplot construction is provided in Section 3. Our visualization tool of the surface boxplot is presented in Section 4. Two applications to samples of brain images and climate model outputs are illustrated in Section 5. The paper ends with a discussion in Section 6.

2 Ranking surfaces

Data depth is an important concept for multivariate data ordering. The general idea is that one can compute the data depth of all the observations and order them according to decreasing depth values. Let $\mathbf{Y}_{[i]}$ denote the observation in $\mathbb{R}^d$ associated with the ith largest depth value. The order statistics, $\mathbf{Y}_{[1]}, \ldots, \mathbf{Y}_{[n]}$, induced by data depth start from the most central data point and move outwards in all directions. The implication is that a smaller rank is associated with a more central position with respect to the data cloud. With regard to functional data, López-Pintado & Romo (2009) introduced the band depth (BD) concept to order sample curves, when each observation is a real function, $y_i(t),\ i = 1, \ldots, n,\ t \in \mathcal{I}$, where $\mathcal{I}$ is an interval in $\mathbb{R}$. According to the general idea of data depth, for sample curves, $y_{[1]}(t)$ is the deepest (most central) curve or simply the median curve, and $y_{[n]}(t)$ is the most outlying curve.

More specifically, López-Pintado & Romo (2009) defined the BD through a graph-based approach. Let the graph of a function, $y(t)$, be the subset of the plane $G(y) = \{(t, y(t)) : t \in \mathcal{I}\}$. Then, the band in $\mathbb{R}^2$ delimited by the curves $y_{i_1}, \ldots, y_{i_k}$ is defined as $B(y_{i_1}, \ldots, y_{i_k}) = \{(t, x(t)) : t \in \mathcal{I}, \min_{r=1,\ldots,k} y_{i_r}(t) \leq x(t) \leq \max_{r=1,\ldots,k} y_{i_r}(t)\}$. Let J be the number of curves determining a band, where J is a fixed value with $2 \leq J \leq n$. If $Y_1(t), \ldots, Y_n(t)$ are independent copies of the stochastic process $Y(t)$ generating the observations $y_1(t), \ldots, y_n(t)$, the population version of the BD for a given curve, $y(t)$, with respect to the probability measure, P, is defined as $BD_J(y, P) = \sum_{j=2}^{J} BD^{(j)}(y, P) = \sum_{j=2}^{J} P\{G(y) \subset B(Y_1, \ldots, Y_j)\}$, where $B(Y_1, \ldots, Y_j)$ is a band delimited by j random curves. The sample version of $BD^{(j)}(y, P)$ is defined as $BD_n^{(j)}(y) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq n} I\{G(y) \subseteq B(y_{i_1}, \ldots, y_{i_j})\}$, where $I\{\cdot\}$ denotes the indicator function. Then, the sample BD of a curve, $y(t)$, is $BD_{n,J}(y) = \sum_{j=2}^{J} BD_n^{(j)}(y)$. The indicator function in the BD definition accounts only for bands that completely contain a sample curve. Hence, the depth values tend to have too many ties, especially when curves are very irregular, such that few bands completely contain a curve. To solve this problem, López-Pintado & Romo (2009) proposed a modified BD (MBD) that replaces the indicator function with a function that measures the proportion of time that a curve, $y(t)$, is in a band. It yields a more flexible ordering of the curves in the sample.

The BD or MBD requires constructing all the possible bands, and the computational cost grows with the sample size n at the rate $\binom{n}{j},\ 2 \leq j \leq J$. López-Pintado & Romo (2009) pointed out that although the number, j, of curves

(wileyonlinelibrary.com) DOI: 10.1002/sta4.39

determining a band could be any integer between 2 and J, the order of curves induced by the BD is very stable in J. To avoid computational issues, $J = 2$ is used by Sun & Genton (2011, 2012a), and a fast BD computation algorithm has been developed by Sun et al. (2012).

Now, suppose we observe sample surfaces, $z_1(\mathbf{s}), \ldots, z_n(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, where $\mathcal{S}$ is a region in $\mathbb{R}^2$. The information unit for such a dataset is the entire surface. To order sample surfaces, we therefore need to generalize univariate order statistics to surfaces. To this end, we generalize the MBD with $J = 2$ to $\mathbb{R}^3$ through a volume. We define the sample modified volume depth (MVD) to be

$$MVD_n(z) = \binom{n}{2}^{-1} \sum_{1 \le i_1 < i_2 \le n} \lambda_r \left\{ A\left(z; z_{i_1}, z_{i_2}\right) \right\},$$

where $A\left(z; z_{i_1}, z_{i_2}\right) \equiv \left\{ \mathbf{s} \in \mathcal{S} : \min_{r=i_1,i_2} z_r(\mathbf{s}) \le z(\mathbf{s}) \le \max_{r=i_1,i_2} z_r(\mathbf{s}) \right\}$ and $\lambda_r(z) = \lambda\left(A\left(z; z_{i_1}, z_{i_2}\right)\right)/\lambda(\mathcal{S})$, if λ is the Lebesgue measure on $\mathbb{R}^3$. A sample median surface is a surface from the sample with the largest sample modified volume depth value, defined by $\arg\max_{z \in \{z_1,\ldots,z_n\}} MVD_n(z)$. If there are ties, the median will be the average of the surfaces maximizing the sample modified volume depth.

3 Surface boxplot construction

The construction of surface boxplots is a strong analogue to that of functional boxplots (Sun & Genton, 2011). The first step is the surface ordering. Sample surfaces are ordered from the centre outwards based on their MVD values, inducing the order $z_{[1]}, \ldots, z_{[n]}$. The sample α central region is naturally defined as the volume delimited by the α proportion $(0 < \alpha < 1)$ of the deepest surfaces. In particular, the sample 50% central region is

$$C_{0.5} = \left\{ (\mathbf{s}, z(\mathbf{s})) : \min_{r=1,\ldots,\lceil n/2 \rceil} z_{[r]}(\mathbf{s}) \le z(\mathbf{s}) \le \max_{r=1,\ldots,\lceil n/2 \rceil} z_{[r]}(\mathbf{s}) \right\},$$

where $\lceil n/2 \rceil$ is the smallest integer not less than $n/2$. The border of the 50% central region is defined as the inner envelope representing the box in a surface boxplot. The median surface in the box is the one with the largest depth value.

Because the ordering is from the centre outwards, the volume of the central region increases as α increases. Hence, the maximum envelope, or the outer envelope, is defined as the border of the maximum non-outlying central region. To determine this region, we propose to identify outlying surfaces by an empirical rule similar to the 1.5 times the 50% central region rule in a functional boxplot. The fences are obtained by inflating the inner envelope by 1.5 times the range of the 50% central region. Any surfaces crossing the fences are flagged as potential outliers. The factor 1.5 can be also adjusted as in the adjusted functional boxplots (Sun & Genton, 2012a) to take into account spatial autocorrelation and possible correlations between surfaces.

4 Visualization

We have created an interactive visualization tool for exploring volumetric slice-based datasets using the surface boxplot to extract descriptive statistics including the median, inner and outer envelopes, and potential outliers. As shown in Figure 1, the visualization tool uses a multi-window approach, coordinating a collection of distinct views via mouse interactions, each aimed at allowing the user to see the data from a unique perspective.

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.39

The ISI's Journal for the Rapid
Dissemination of Statistics Research

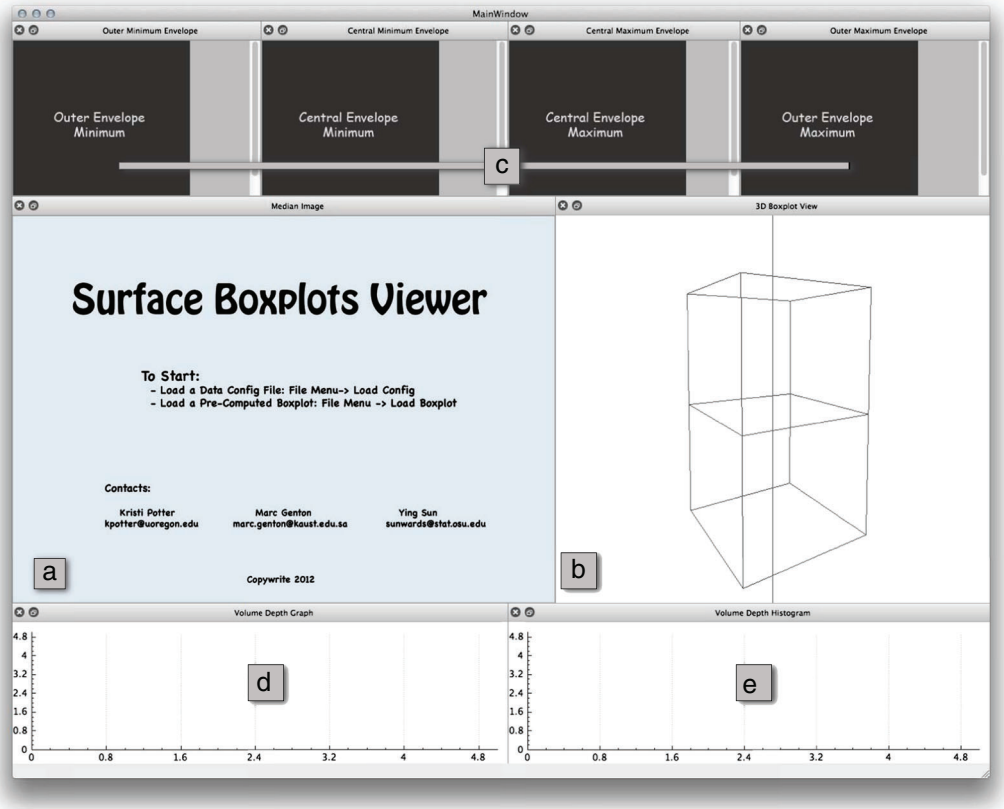

Figure 1. Overview of the surface boxplot visualization tool: (a) the median display; (b) the 3D boxplot display; (c) the envelope display; (d) the volume depth graph; (e) the volume depth histogram.

4.1. Median display

At the centre of the display (Figure 1a) is the median display. This display shows the median surface from the dataset, which is the middlemost surface and can be thought of as a representative of the data. The display allows the user to zoom in and scroll around the image to allow for in-depth and contextual views. We have chosen this display as the largest, centralized display because it will be used as a comparison image throughout the exploration of the dataset.

4.2. 3D boxplot display

Also in the centre is the 3D boxplot display (Figure 1b). This display encodes the median and envelope images as heightfields to allow a quick comparison between all images. The median image is displayed as the central heightfield, and minimum and maximum images from the inner and outer envelopes are displayed above and below the median image, respectively. Figure 2 shows a close-up of a 3D boxplot. The user can rotate, zoom, and pan the boxplot to gain a better understanding and change the colour of the background for a better display of the data.

4.3. Envelope display

At the top of the tool, Figure 1c, are the displays of the inner and outer envelopes. From left to right is the minimum outer, minimum central, maximum central, and maximum outer envelope images. These images are composited pixel-

(wileyonlinelibrary.com) DOI: 10.1002/sta4.39

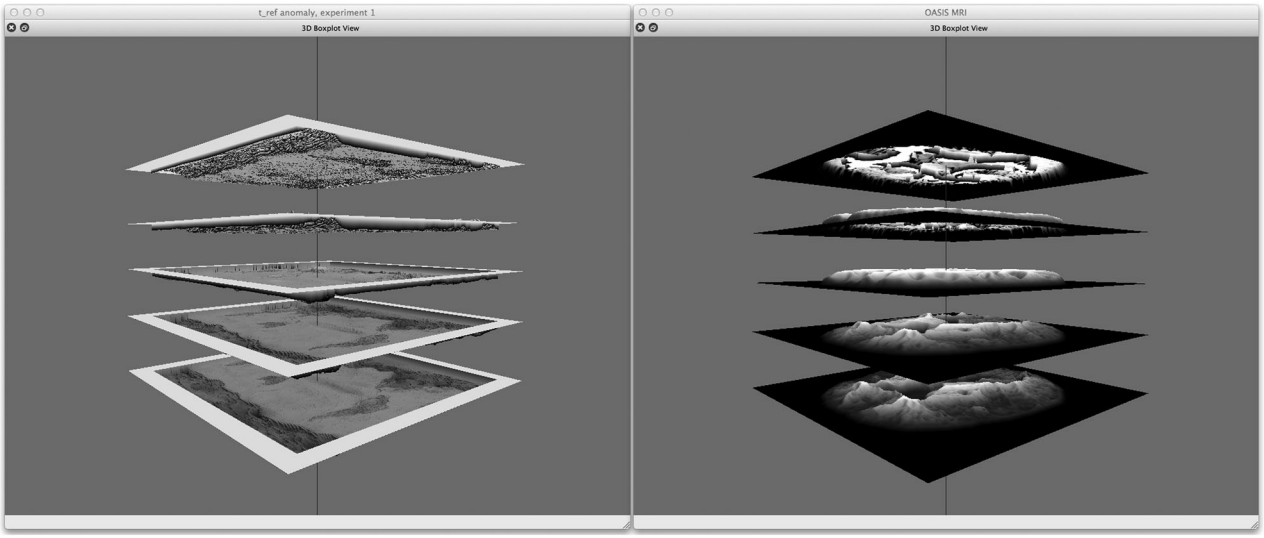

Figure 2. Two examples of the 3D boxplot display. Each image is encoded as a heightfield. The median surface is the central heightfield, flanked by the inner and outer envelopes.

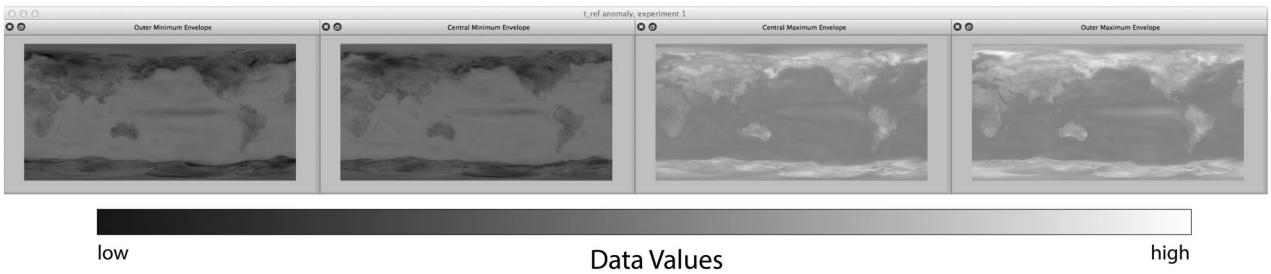

Figure 3. Envelope display. The outer envelope is shown on the far left (minimum) and far right (maximum), while the central envelope is shown in the two centre images.

wise from the entire dataset. They are not actual data realizations and thus displayed in greyscale even for original colour images. The interpretation of these images is an indication of the overall minimal and maximal pixel values (not including potential outliers) and minimum and maximum pixel values of the central 50% of the data, both of which create envelopes that can be thought of like the quartiles making up the "box" of a traditional 1D boxplot. Figure 3 shows an example of the envelope display using temperature data from a climate modelling simulation. As shown in the figure, dark pixels represent low values, while light pixels represent high values.

4.4. Volume depth graph and histogram

To understand the results from calculating the surface boxplot, we have added a volume depth graph (Figures 1d and 4). This graph plots the volume depth of every image in the dataset such that the index number of the image is on the *x*-axis and the volume depth is on the *y*-axis. The median image is indicated by a solid red disc; potential outliers are shown as blue stars; and all other images are outlined black discs. This display allows the user to see the number of potential outliers that exist in the data, as well as the volume depth of those outliers.

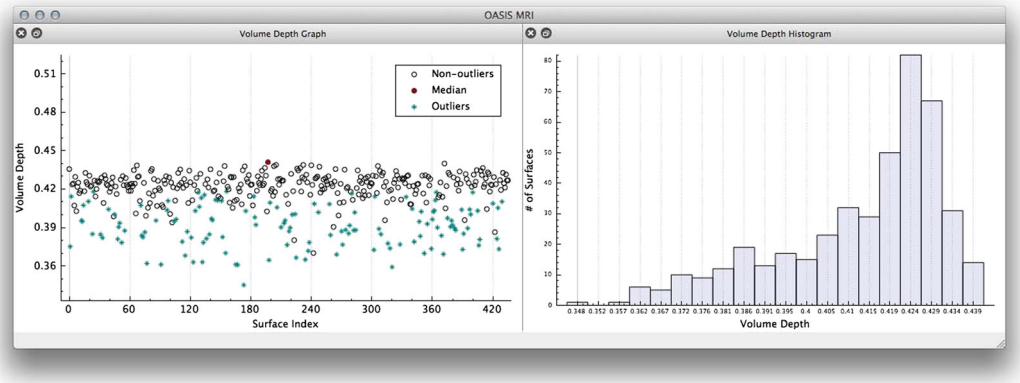

Figure 4. A volume depth graph (left) and a volume depth histogram (right).

A volume depth histogram is also included in the tool (Figures 1e and 4), which summarizes the volume depth across the entire dataset. The graph plots the number of images in each histogram bin, allowing the user to see quickly the volume depth range containing the largest and smallest numbers of images.

4.5. Interactions

The tool is designed to be highly interactive to allow for exploration and comparison. Every image display can be pulled out from the tool into its own separate window and placed anywhere on the screen. All images have zoom and scrolling functionality through scroll bars and keyboard interactions. The volume depth graph allows the user to zoom in and pan out in the graph itself to ensure that all data are viewable, or for more in-depth investigation. Multiple points can be selected by the user, which are then highlighted, and upon a shift-modified click, a new independent image display pops up with the corresponding image, as shown in Figure 5. This new image display also allows for scrolling and zooming, and it can be placed anywhere within the screen for comparisons. Similarly, the volume depth histogram allows for the selection of bins via the mouse, as shown in Figure 6. Upon selection, the bin is highlighted in blue, and the corresponding images in the volume depth graph are also highlighted. A shift-modified mouse click will bring up all images within the bin in independent image displays. These interactions allow the user to investigate single images or entire ranges of images. For example, selecting histogram bins corresponding to outliers allows the user to bring up all outliers quickly for investigation.

4.6. Implementation

The application is developed using C++ and Qt. While the application was developed to explore our application-specific images, it is flexible enough to work on any collection of image data. The code is available in the supporting information of this paper.

5 Applications

5.1. Open Access Series of Imaging Studies brain images

The first application of the surface boxplot is the Open Access Series of Imaging Studies (OASIS) brain magnetic resonance imaging dataset (Marcus et al., 2007). This dataset consists of a collection of 436 brain slices of subjects aged

(wileyonlinelibrary.com) DOI: 10.1002/sta4.39

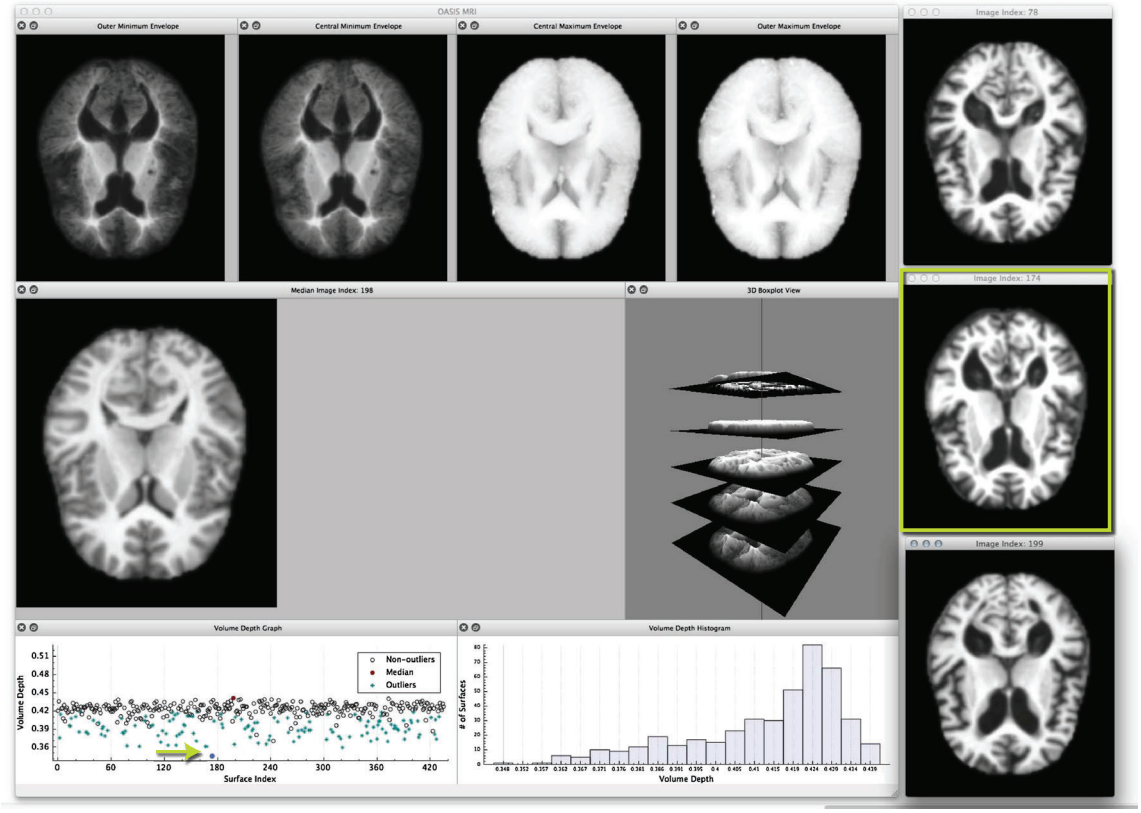

Figure 5. An example of the tool's reaction when a point on the graph is selected. When the user selects a point, it is highlighted, and a shift-modified selection will create an independent image display of the corresponding image.

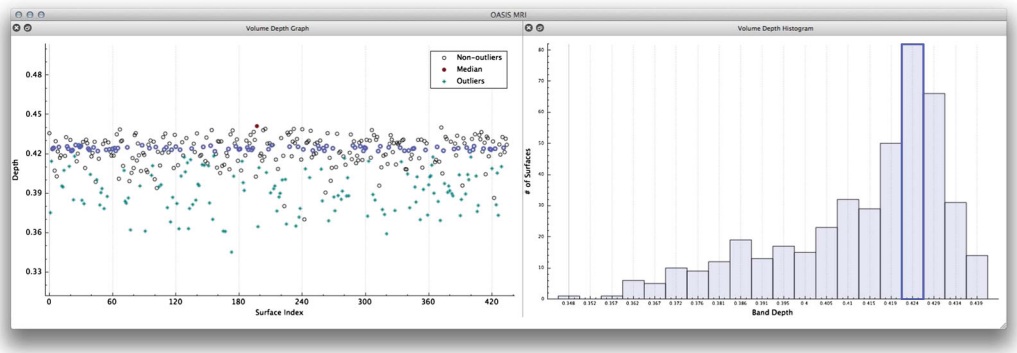

Figure 6. An example of the tool's reaction when a bin on the histogram is selected. When the user selects a bin, it is highlighted along with all corresponding images in the volume depth graph. Analogous to the volume depth selection system, a shift-modified click will bring up all images within the bin in their own independent image displays.

18–96 years and includes a subset of subjects who have been diagnosed with very mild to moderate Alzheimer's disease. We have applied the surface boxplot to this dataset to try to determine non-normal brain functioning by

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.39

The ISI's Journal for the Rapid
Dissemination of Statistics Research

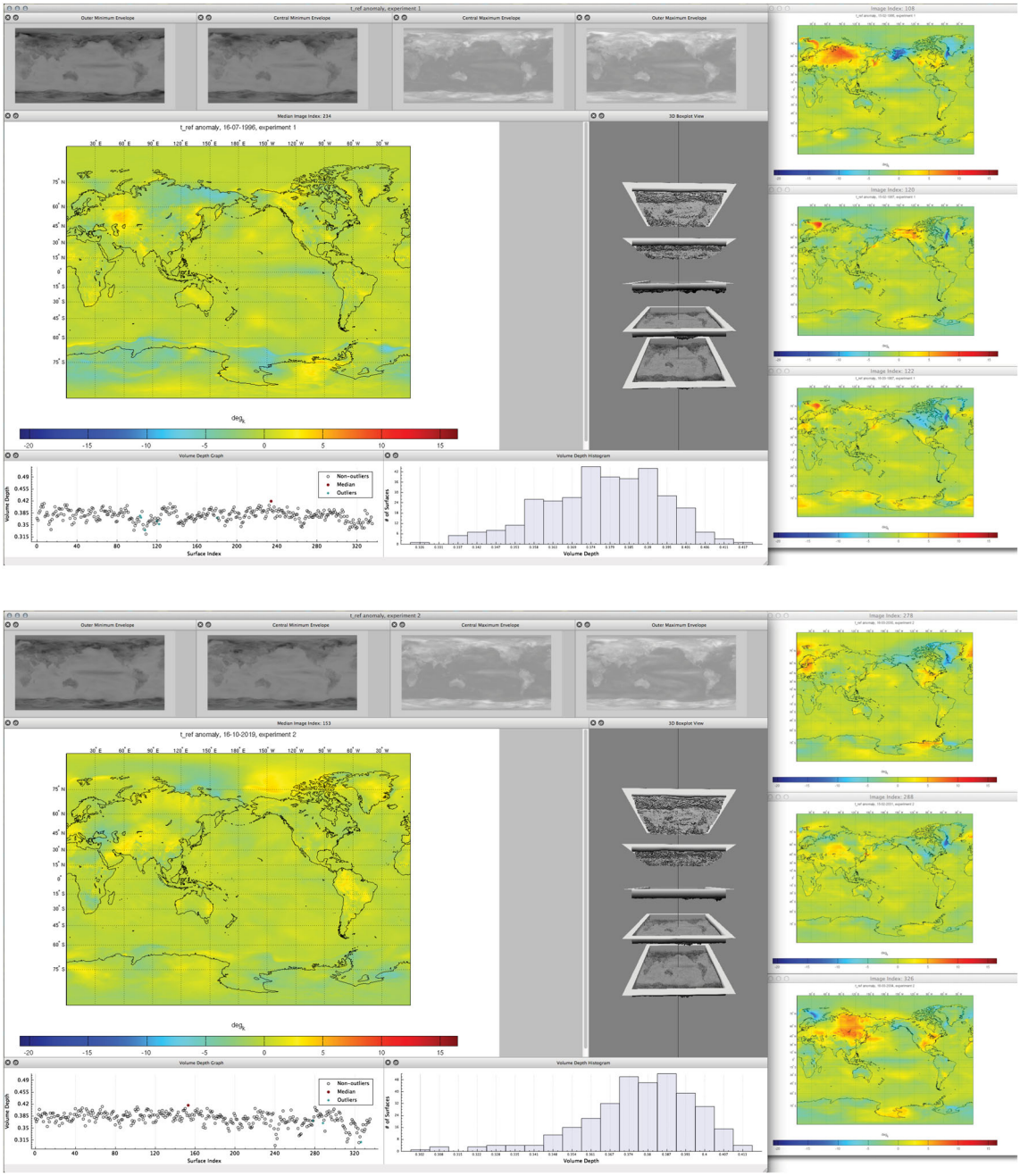

Figure 7. Climate model simulation results visualized using our surface boxplot.

identifying scans that lie outside the range of normal brains. That is, we wish to identify potential outliers in the data. Figure 5 shows the results of applying the surface boxplot on these data. We can see, on the right of Figure 5, three brain images with low volume depth values that have been identified as potential outliers in this dataset. These images have depth values of 0.3619, 0.3451, and 0.3645, and they have clear visual differences in comparison with the

The ISI's Journal for the Rapid
Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.39

median image that has a depth value of 0.4410. Because of the complexity of medical diagnoses, we cannot say directly that the potential outlier images found using our technique identify Alzheimer's patients, although we are able to select brain images outside the range of normal that may, with further testing, indicate some type of impairment.

5.2. Climate model outputs

We have also tested our method on outputs of climate model simulations from the high-resolution atmospheric model (HIRAM) at the National Oceanic and Atmospheric Administration Geophysical Fluid Dynamics Laboratory (GFDL) in Princeton (Zhao et al., 2009). The simulations were conducted on the cubed sphere with an effective resolution of 25 km. The data cover the time period 1977–2004 (experiment 1, historical run) and 2007–2034 (experiment 2, future projection with medium–low Representative Concentration Pathway, RCP4.5) at a monthly resolution and are described by image anomalies of air surface temperature (at 2 m in °K). That is, each data value has been centred with respect to its month's average. The samples from both experiments have $n = 336$ images that are 1648×826 pixels in dimension. The runs have sea surface temperatures (SST) taken from the GFDL earth system model (ESM2M).

Figure 7 presents two surface boxplots, one for each of the experiments. The surface index in the volume depth graphs corresponds to months of each period in increasing order. The rankings of the images reveal interesting features. For example, we see different spatial temperature patterns between the current median from experiment 1 (07/1996) and the future median from experiment 2 (10/2019). Interestingly, the most representative image of experiment 1 is in July, whereas it is in October for experiment 2. The histograms of the volume depth values indicate a more left-skewed shape for experiment 2, that is, more unusual images for the future projection. Three outlying images have been selected to the right of each surface boxplot. For experiment 1, the outliers are for the dates 02/1986, 02/1987, and 03/1987, whereas for experiment 2, they are for the dates 03/2030, 02/2031, and 03/2034. For both experiments, those outliers are in February and March, but intriguingly, for experiment 2, they are at the end of the period. Notice that the outlying images of both experiments clearly show spatial regions of cooling or warming compared to the median images. For experiment 1, the outliers are observed from 1986 to 1987, which are two El Niño years. Other years with local minimum depth values in the volume depth graph are also associated with the El Niño effect, indicating relatively unusual temperature behaviour. For experiment 2, although it produces more unusual images, this effect is not clear in terms of the volume depth values.

6 Discussion

This paper proposed the surface boxplot as a tool for visualization and exploratory analysis of samples of images. We used the notion of volume depth, a generalization of BD, to order the images viewed as surfaces. In particular, we defined the median image of the sample. We used an exact and fast algorithm for the ranking of the images. This allowed us to detect outlying images that often contain interesting features not present in most of the images.

We built a graphical tool to visualize the surface boxplot and its various characteristics. A graph and histogram of the volume depth values allow us to identify images of interest. The code is available in the supporting information of this paper. We applied our surface boxplot to a sample of brain images and to a sample of climate model outputs and then identified various interesting images from these datasets.

An extension of our surface boxplot to multivariate images, that is, to images of more than one variable, could be explored by ranking the images with the simplicial BD introduced by López-Pintado et al. (2014).

Acknowledgement

The authors thank Sergey Osipov at King Abdullah University of Science and Technology (KAUST) for formatting the climate model output image data. This work was supported in part by award no. KUS-C1-016-04 made by KAUST.

References

Blackwell, M, Nikou, C, DiGioia, AM & Kanade, T (2000), 'An image overlay system for medical data visualization', *Medical Image Analysis*, **4**, 67–72.

Jornsten, R (2004), 'Clustering and classification via the L_1 data depth', *Journal of Multivariate Analysis*, **90**, 67–89.

Li, J & Liu, R (2004), 'New nonparametric tests of multivariate locations and scales using data depth', *Statistical Science*, **19**, 686–696.

López-Pintado, S & Jornsten, R (2007), 'Functional analysis via extensions of the band depth', *IMS Lecture Notes-Monograph Series*, IMS **54**, 103–120.

López-Pintado, S & Romo, J (2009), 'On the concept of depth for functional data', *Journal of the American Statistical Association*, **104**, 718–734.

López-Pintado, S, Sun, Y, Lin, J & Genton, MG (2014), 'Simplicial band depth for multivariate functional data', *Advances in Data Analysis and Classification,* to appear.

McAuliffe, M, Lalonde, FM, McGarry, D, Gandler, W, Csaky, K & Trus, BL (2001), 'Medical image processing, analysis and visualization in clinical research', in *Proceedings of the 14th IEEE Symposium on Computer-based Medical Systems (CBMS2001)*, *IEEE Computer Society*, Los Alamitos, CA, 381–386.

Marcus, DS, Wang, TH, Parker, J, Csernansky, JG, Morris, JC & Buckner, RL (2007), 'Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults', *Journal of Cognitive Neuroscience*, **19**, 1498–1507.

Sun, Y & Genton, MG (2011), 'Functional boxplots', *Journal of Computational and Graphical Statistics*, **20**, 313–334.

Sun, Y & Genton, MG (2012a), 'Adjusted functional boxplots for spatio-temporal data visualization and outlier detection', *Environmetrics*, **23**, 54–64.

Sun, Y & Genton, MG (2012b), 'Functional median polish', *Journal of Agricultural, Biological, and Environmental Statistics*, **17**, 354–376.

Sun, Y, Genton, MG & Nychka, D (2012), 'Exact fast computation of band depth for large functional datasets: how quickly can one million curves be ranked?', *Stat*, **1**, 68–74.

Walter, T, Shattuck, DW, Baldock, R, Bastin, ME, Carpenter, AE, Duce, S, Ellenberg, J, Fraser, A, Hamilton, N, Pieper, S, Ragan, MA, Schneider, JE, Tomancak, P & Hériche, JK (2010), 'Visualization of image data from cells to organisms', *Nature Methods*, **7**, S26–S41.

Zhao, M, Held, IM, Lin, S-J & Vecchi, JA (2009), 'Simulations of global hurricane climatology, interannual variability, and response to global warming using a 50 km resolution GCM', *Journal of Climate*, **33**, 6653–6678.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.39

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web site.