Statistical Atlas Construction via Weighted Functional Boxplots

Yi Hong[a], Brad Davis[c], J.S. Marron[a], Roland Kwitt[e], Nikhil Singh[a], Julia S. Kimbell[a], Elizabeth Pitkin[a], Richard Superfine[a], Stephanie D. Davis[d], Carlton J. Zdanski[a], Marc Niethammer[a,b]

[a] *University of North Carolina (UNC) at Chapel Hill, NC, US*
[b] *Biomedical Research Imaging Center, UNC-Chapel Hill, NC, US*
[c] *Kitware, Inc., Carrboro, NC, US*
[d] *Indiana University School of Medicine, Indianapolis, IN, US*
[e] *Department of Computer Science, University of Salzburg, Austria*

Abstract

Atlas-building from population data is widely used in medical imaging. However, the emphasis of atlas-building approaches is typically to estimate a spatial alignment to compute a mean / median shape or image based on population data. In this work, we focus on the statistical characterization of the population data, once spatial alignment has been achieved. We introduce and propose the use of the *weighted* functional boxplot. This allows the generalization of concepts such as the median, percentiles, or outliers to spaces where the data objects are functions, shapes, or images, and allows spatio-temporal atlas-building based on kernel regression. In our experiments, we demonstrate the utility of the approach to construct statistical atlases for pediatric upper airways and corpora callosa revealing their growth patterns. We also define a score system based on the pediatric airway atlas to quantitatively measure the severity of subglottic stenosis (SGS) in the airway. This scoring allows the classification of pre- and post-surgery SGS subjects and radiographically normal controls. Experimental results show the utility of atlas information to assess the effect of airway surgery in children.

Keywords: Statistical atlas-building, weighted functional boxplots, kernel regression, pediatric upper airways.

1. Introduction

Atlas-building from population data has become an important task in medical imaging to provide templates for data analysis. Numerous methods for atlas-building exist, ranging from methods designed for cross-sectional, longitudinal, and random design data. These approaches typically estimate a representative data object (Wang and Marron, 2007) (e.g., shape, surface, image) for the population; e.g., a population mean (Joshi et al., 2004) or median (Fletcher et al., 2009) with respect to spatial deformations and appearance.

A limitation of all these methods is that they result in a single summary representer and discard much of the population for subsequent analysis. For instance, a single point is used to summarize the entire population on the manifold, when one summarizes it using an atlas or a median. For regression, a single curve summarizes the population without carrying forward any information from the local distribution of data around the curve. These are restrictive representations that limit the capability of the model to present confidence bounds, quantile measurements or to identify outliers. In the literature, the limitation of the single summary representers has also been acknowledged. For instance, Aljabar et al. (2009) suggest a multi-atlas approach to estimate multiple representers of the population. In another study, Gerber et al. (2010) propose to learn a low-dimensional representation driven entirely by the population of images.

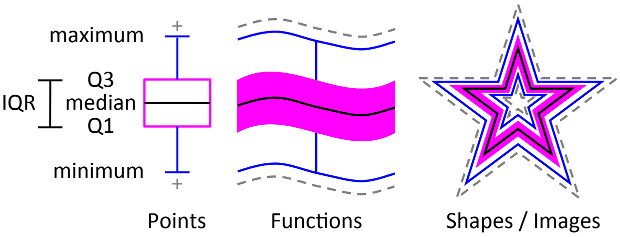


Figure 1: Illustration of boxplots for points, functions, shapes and images. Median (middle black line), confidence region (magenta) and the maximum non-outlying envelope (two outward blue lines). The gray dashed lines are the outliers.

Another strategy to retain population variation information is to represent additional aspects of the full data distribution, such as percentiles, the minimum and maximum, variance, confidence regions and outliers as captured by a boxplot for scalar-valued data. The functional boxplot (Sun and Genton, 2011) is an effective tool to represent such statistics for functions. The main goal in this paper is to generalize the notion of functional boxplots to summarize variabilities within population of entities such as shapes and images (see Fig. 1). This can provide a simple and generic method to augment atlases with additional population information while avoiding restrictive point-wise analyses of data-objects. Note that we focus in this paper on augmenting atlases with statistical information, and assume a given spatial alignment of data objects. The method can also be extended to build order statistics from low-dimensional manifold embeddings where point-wise analysis becomes meaningful as each point then represents a full data object.

Another goal of this work is to generalize the notion of confidence bounds to the estimates of regression using functional boxplots. As subject data typically has associated individual characteristics (e.g., age, weight, gender) we want to be able to compute the statistical information parameterized by these characteristics. For example, given a subject at a particular age we want to compute subject age-specific confidence regions to assess similarity with respect to the full data population.

We make the following contributions in this paper:

- *We develop a weighted variant of the functional boxplot* in Section 2. This enables for example the use of kernel-regression to build spatio-temporal atlases.
- *We show the effectiveness of the method in comparison to point-wise analysis* in Section 4 highlighting the importance of object-oriented data analysis.
- *We show applicability of the method to functions, shapes, and images* in Section 6 and demonstrate how an atlas can robustly be augmented with statistical data for two applications: capturing changes in pediatric airway development and changes of the corpus callosum over time. We also briefly sketch how our method could be used to build order-statistics on manifolds.
- *We show the use of our method for quantitative assessment in pediatric airways* in Section 8, where an age-adapted atlas can be used to score the severity of a child suffering from airway obstruction before and after surgery. This quantitative assessment shows significant differences among normal controls, pre- and post-surgery SGS subjects.

The method described in this paper is an extension of the preliminary ideas we presented in a recent conference paper (Hong et al., 2013a). This paper offers more details of our proposed method, additional experiments for quantitative assessment, and more validation on synthetic and real data.

2. Weighted functional boxplots and atlas-building

This section introduces a weighted variant of the functional boxplot and extends it for use with kernel regression of functional data. We first cover the preliminaries on kernel regression and later present the concept of weighted band-depth essential to defining the weighted functional boxplot. The proposed method is applicable to the analysis of function, shape, and image populations to create non-parametric regression models with associated subject characteristics. As an example, we consider subject age and demonstrate the effectiveness of weighted functional boxplots and kernel smoothing (Wand and Jones, 1994) to build a spatio-temporal atlas.

2.1. Atlas building with kernel regression

Given spatially aligned data objects we want to capture population changes, e.g., with respect to age. Spatial alignment refers to a pre-processing step that transforms all data objects to common coordinates for further analysis. The type of alignment depends on the objectives of a particular study. For instance, this alignment may be a rigid transformations when the statistical analysis needs to be performed modulo translations and rotations only. An atlas with population changes can be built through kernel regression which assigns weights to data-objects with respect to the regressor (say a desired age $\bar{a}$). For example, we can use a Gaussian weighting function $w_i(a_i; \sigma, \bar{a}) = ce^{-(a_i-\bar{a})^2/2\sigma^2}$, where a_i is the age of the observation i, σ is the Gaussian standard deviation and c a normalization constant to assure that the weights sum up to one.

2.1.1. Boundary bias

Kernel-based methods exhibit a bias near the boundary of the available data. This is usually attributed to the asymmetric averaging of limited information at the boundaries. Many solutions have been proposed to address this issue (Schuster, 1985; Jones, 1993; Marron and Ruppert, 1994). If the target age for the atlas is located within the interior part of the observed population, no boundary effects exist. However, for studies involving models for growth, aging or memory decline, we often build atlases for very young or very old subjects. This usually requires averaging kernel weights with respect to an age near the boundary. In such models, to mitigate the boundary bias, we adjust the weights around the boundaries based on the approach proposed in Schuster (1985), which relies on adjustment through boundary reflection.

We assume observations are given in the age range $[b_l, b_h]$. We adjust weights for observations at the boundaries in kernel regression by folding using reflection. In particular, given the kernel bandwidth, σ, and the location for each observation, a_i, with respect to the regression location, $\bar{a}$, the adjusted weights over the complete range are given as

$$w_i(a_i; \sigma, \bar{a}) = \begin{cases} c(g(a_i) + g(2b_l - a_i)), & a_i \in [b_l, b_l + \sigma) \\ cg(a_i), & a_i \in [b_l + \sigma, b_h - \sigma] \\ c(g(a_i) + g(2b_h - a_i)), & a_i \in (b_h - \sigma, b_h]. \end{cases} \tag{1}$$

Here $g(\cdot)$ denotes the Gaussian function, $g(\cdot; \sigma, \bar{a})$, with the mean, $\bar{a}$, and the variance, σ^2, as mentioned above.

2.1.2. Bandwidth for kernel

An appropriate choice of the bandwidth, σ, for kernel regression depends upon the application. In general, the bandwidth should be chosen based on the expected variation in the data. A small bandwidth is able to express fast changes at the potential cost of becoming noise-sensitive, whereas a bandwidth that is too large gives overly smooth kernel regression results. A compromise can be achieved by selecting the bandwidth through cross-validation based model selection procedures (Kohavi, 1995). We will cover more details on how to choose σ in the experiment sections.

Note that for scalar-valued data, the weights presented in Section 2.1.1 can be used to define a weighted mean. For kernel regression on deformations, these weights can be incorporated during the atlas-building procedure, e.g., in atlas construction using images (Davis et al., 2010). Our main goal is to augment the atlas with statistical information about observations and hence develop a weighted functional boxplot. This will allow us to obtain a regressed median. The median will be one of the data-objects from the population which represents the center at the target age. We will further define the α central region and compute the interquartile-range, the maximum non-outlying envelope, and detect outliers.

2.2. Weighted functional boxplots

The challenge in defining a *functional* boxplot is to develop a notion of ordering for the space of functions. Once this ordering has been defined order statistics can be computed. Hence, the equivalent to a scalar-valued boxplot which makes use of the median and percentiles of the data can be defined. Sun et al. (Sun and Genton, 2011) proposed an ordering of functions based on the concept of band-depth. Essentially, band-depth measures how deeply a particular function is buried within all the other functions of the data population. The deepest one is then declared the median curve. The band-depth itself is used to define the ordering among functions.

To define a *weighted* functional boxplot consistent with the functional boxplot introduced by Sun et al. (Sun and Genton, 2011) requires the definition of a consistent *weighted band depth* for functional data.

2.2.1. Weighted band-depth

The functional boxplot is defined through the concept of band-depth (López-Pintado and Romo, 2009; Sun and Genton, 2011). Since in our case, each observation has a different weight, we first need to define a weighted band-depth. Such a definition would naturally generalize the functional boxplot to the weighted functional boxplot.

To motivate our choice, assume we want to compute a standard weighted median of scalar values. This is given by

$$\mu^* = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} w_i |x_i - \mu|, \tag{2}$$

where μ is the sought-for median, n is the number of measurements, $\{x_i\}$ are the measurements, and $w_i > 0$ are weights for the individual measurements. Assume that all weights are natural numbers, i.e., $w_i \in \mathbb{N}^+$. This can be realized exactly for arbitrary rational, w_i, and in general by multiplying the energy with a suitable constant, which does not change the minimizer. Hence, we replace the weighted problem with the equivalent unweighted minimization problem

$$\mu^* = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} |x_i - \mu|, \tag{3}$$

where the individual measurements are simply repeated based on their multiplicities, $m_i = w_i$.

Using a similar strategy, we can derive the weighted band-depth. The band-depth introduced in (López-Pintado and Romo, 2009; Sun and Genton, 2011) is defined for a population of n functions, y_i (for $i = 1 \ldots n$), defined on a domain $\mathbb{I}$, where $\mathbb{I}$ is an interval in $\mathbb{R}$. It is a graph-based approach that computes the fraction of bands delimited by the subset of the population containing the curve, $\mathrm{y}(\mathbf{x})$. In particular, it is defined as

$$BD_n^{(j)}(\mathrm{y}) = \frac{1}{C} \sum_{1\le i_1<i_2<\cdots<i_j\le n} I\{G(\mathrm{y}) \subseteq B(\mathrm{y}_{i_1}, \cdots, \mathrm{y}_{i_j})\}. \tag{4}$$

Here j is the number of observations used for defining the band, C is a normalization constant equal to the number of admissible permutations. $G(\mathrm{y})$ is the graph of the function, $G(\mathrm{y}) = \{(\mathbf{x}, \mathrm{y}(\mathbf{x})) : \mathbf{x} \in \mathbb{I}\}$. B is the band delimited by the observations given as its arguments. That is, $B(\mathrm{y}_{i_1}, \cdots, \mathrm{y}_{i_j}) = \{(\mathbf{x}, \mathrm{y}(\mathbf{x})) : \mathbf{x} \in \mathbb{I}, min_{r=i_1,\cdots,i_j} \mathrm{y}_r(\mathbf{x}) \le \mathrm{y}(\mathbf{x}) \le max_{r=i_1,\cdots,i_j} \mathrm{y}_r(\mathbf{x})\}$. $I\{.\}$ denotes the indicator function, which evaluates to 1 if the graph of the function is within the band or to 0 otherwise.

Now we want to define a weighted variant of the above definition of band-depth. For the weighted variant, say, we are now given a population of functions, y_i, for $i = 1 \ldots n$, each with its associated weight, w_i. Before we present the actual expression for weighted band-depth, we first write its repeated version. We notice that, similar to the scalar case, we could write the band-depth for this population of functions by repeating each observation as per its given weight. The band-depth with repeats is then given as

$$BD_{\overline{n}}^{(j)}(\mathrm{y}) = \frac{1}{C'} \sum_{1\le i_1<i_2<\cdots<i_j\le \overline{n}} I\{G(\mathrm{y}) \subseteq B(\overline{\mathrm{y}}_{i_1}, \cdots, \overline{\mathrm{y}}_{i_j})\},$$

$$\text{s.t. } \{\overline{\mathrm{y}}_{i_1}, \cdots, \overline{\mathrm{y}}_{i_j}\} \text{ contains unique observations,} \tag{5}$$

where C' is the normalization constant representing admissible permutations adjusted for repeats and $\overline{n}$ is the number of observations including the repeats. The $\{\overline{\mathrm{y}}_i\}$ contain the original observations $\{\mathrm{y}_i\}$, but with repeats, according to their respective multiplicity given by their weights. The band with repeated observations is given as $B(\overline{\mathrm{y}}_{i_1}, \cdots, \overline{\mathrm{y}}_{i_j}) = \{(\mathbf{x}, \mathrm{y}(\mathbf{x})) : \mathbf{x} \in \mathbb{I}, min_{r=i_1,\cdots,i_j} \overline{\mathrm{y}}_r(\mathbf{x}) \le \mathrm{y}(\mathbf{x}) \le max_{r=i_1,\cdots,i_j} \overline{\mathrm{y}}_r(\mathbf{x})\}$. We made use of the fact that, according to our definition, we only want to consider unique observations for the depth measure.

Finally, we define the weighted band-depth by rewriting the sampled band-depth as

$$WBD_n^{(j)}(\mathrm{y}) = \frac{1}{\sum_{1\le i_1<i_2<\cdots<i_j\le n} w_{i_1} w_{i_2} \cdots w_{i_j}} \cdot \sum_{1\le i_1<i_2<\cdots<i_j\le n} w_{i_1} w_{i_2} \cdots w_{i_j} I\{G(\mathrm{y}) \subseteq B(\mathrm{y}_{i_1}, \cdots, \mathrm{y}_{i_j})\}, \tag{6}$$

which generalizes to non-natural-numbered weights $w_i \in \mathbb{R}^+$. This is a natural way to define a weighted band-depth and, in further consequence, a weighted functional boxplot. Computing the weighted band-depth in this way is intuitive, as only bands with large weights for all its individual observations have a large impact. Furthermore, this weighted version can also be adapted to the modified band-depth proposed in Sun and Genton (2011), i.e.,

$$WMBD_n^{(j)}(\mathrm{y}) = \frac{1}{\sum_{1\le i_1<i_2<...<i_j\le n} w_{i_1}w_{i_2}\cdots w_{i_j}} \cdot \sum_{1\le i_1<i_2<...<i_j\le n} w_{i_1}w_{i_2}\cdots w_{i_j}\lambda_m\{A(\mathrm{y};\mathrm{y}_{i_1},...,\mathrm{y}_{i_j})\} \tag{7}$$

where $A_j(\mathrm{y}) \equiv A(\mathrm{y};\mathrm{y}_{i_1},...,\mathrm{y}_{i_j}) \equiv \{\mathbf{x} \in \mathbb{I} : min_{r=i_1,...,i_j}\mathrm{y}_r(\mathbf{x}) \le \mathrm{y}(\mathbf{x}) \le max_{r=i_1,...,i_j}\mathrm{y}_r(\mathbf{x})\}$, m is the observation's dimension, $\lambda_m(\mathrm{y}) = \lambda(A_j(\mathrm{y}))/\lambda(\mathbb{I})$ and λ is the Lebesgue measure on $\mathbb{R}^m$.

With the above definitions, the band depths of all the sampled observations can be calculated and ranked in descending order, $\mathrm{y}_{[1]}(\mathbf{x}) \ge ... \ge \mathrm{y}_{[n]}(\mathbf{x})$. $\mathrm{y}_{[1]}(\mathbf{x})$ is the deepest observation and regarded as a notion of the median of the population, whereas $\mathrm{y}_{[n]}(\mathbf{x})$ is the "most outlying" observation which is a potential outlier.

2.2.2. *α central region*

The concept of central region was introduced in Liu et al. (1999). We define the α central region for the weighted functional boxplot based on the weights of observations. The band of the α central region is delimited by the α proportion of all weights, i.e., the accumulated weights of the first $\hat{p}$ deepest observations.

We first compute the value of $\hat{p}$ based on the weights by

$$\hat{p} = \{p \in \mathbb{N}^+ : \sum_{r=1,...,p-1} w_{[r]} < \alpha, \sum_{r=1,...,p} w_{[r]} \ge \alpha, and\ p \le n\}, \tag{8}$$

where $w_{[r]}$ corresponds to the weight for the r-th deepest observation and $0 \le \alpha \le 1$. Here, we assume that the weights are normalized so they sum up to one. Then the α central region can be generated using these first $\hat{p}$ observations through

$$WCR_\alpha = \{(\mathbf{x}, \mathrm{y}(\mathbf{x})) : \min_{r=1,...,\hat{p}} \mathrm{y}_{[r]}(\mathbf{x}) \le \mathrm{y}(\mathbf{x}) \le \max_{r=1,...,\hat{p}} \mathrm{y}_{[r]}(\mathbf{x})\}. \tag{9}$$

When $\alpha = 0.5$, Eq. (9) corresponds to the 50% central region $WCR_{0.5}$. In practice, the 50% central region is commonly chosen as the confidence region for analysis because it 1) is a robust range for interpretation and 2) enables visualization of the data spread which is less affected by outliers or extreme-values.

2.2.3. *Outlier detection*

In classical boxplots, the outliers can be detected by the 1.5 IQR (interquartile range) (Frigge et al., 1989). This is comparable to 1.5 times the height of the 50% central region for the weighted functional boxplot. The weights of the observations also need to be taken into consideration during the outlier detection. For a Gaussian distribution, the IQR encompasses the most central 50% of the distribution and the fence defined by 1.5 IQR covers 99.3% of the distribution. Therefore, we use this threshold, 0.993, to find the first $\hat{q}$ deepest observations that would be within the fences by

$$\hat{q} = \{q \in \mathbb{N}^+ : \sum_{r=1,...,q-1} w_{[r]} < \beta, \sum_{r=1,...,q} w_{[r]} \ge \beta, and\ q \le n\}, \tag{10}$$

where $\beta = 0.993$. The next step is to narrow the fences with the 1.5 IQR, so that the fences defined in weighted functional boxplots are the combination of the fence defined by the 1.5 IQR with the accumulated weights consistent with the 1.5 IQR of the normal distribution:

$$\begin{aligned} &C_{fences} = \{(\mathbf{x}, \mathrm{y}(\mathbf{x})) : \\ &max(min_{r=1,...,\hat{q}}\mathrm{y}_{[r]}(\mathbf{x}), min(WCR_{0.5}) - 1.5 * IQR) \cup \\ &min(max_{r=1,...,\hat{q}}\mathrm{y}_{[r]}(\mathbf{x}), max(WCR_{0.5}) + 1.5 * IQR)\}. \end{aligned} \tag{11}$$

Any objects outside the fences defined by C_{fences} are flagged as outliers.

3. Implementation and algorithm complexity

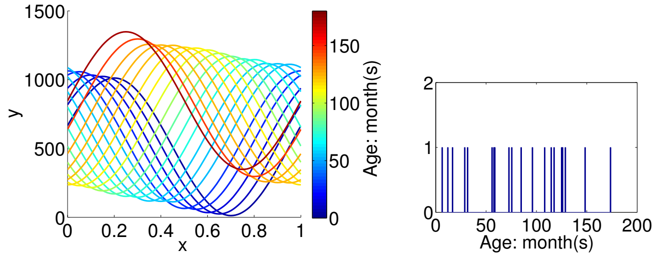
In this section, we discuss how to implement our statistical atlas-building method based on the weighted functional boxplots (see Algorithm 1), as well as the time complexity of the algorithm. From the observations and their associated independent values, e.g., ages, the algorithm generates a statistical atlas at a target age, which consists of the median, the confidence region, the maximal non-outlying region, and the outliers. Details about converting between the functional representation of the boxplot and shapes and images are covered in Section 6.2.

Algorithm 1: Statistical atlas-building based on weighted functional boxplots

```
Data: $\{a_i, \mathrm{y}_i\}_{i=1}^N$ ($N$ observations with ages), $\bar{a}$ (the target
      age), $J$ (the number of observations for a band)
Result: $\mathrm{y}_{[1]}$ (the median), $WCR_\alpha$ (the $\alpha$ center region),
        $C_{fences}$ (the fences of the atlas)
Choose the bandwidth $\sigma$ and compute the weight $w_i$ for
each $\mathrm{y}_i$ with Gaussian function centered at $\bar{a}$ (Section 2.1).
for $i$ := 1:N do
    for $j$ := 2:J do
        Loop through all combinations, choosing j from N
        observations, and compute the weighted band
        depth for $\mathrm{y}_i$ using Eq. 6 or 7.
    end
end
Sort $\{\mathrm{y}_i\}_{i=1}^N$ based on the weighted band-depth in the
decreasing order, $\mathrm{y}_{[1]} \ge \cdots \ge \mathrm{y}_{[N]}$.
Compute $WCR_\alpha$ and $C_{fences}$ according to Section 2.2.2 and
Section 2.2.3 respectively.
```

Since in practice $J \ll n$, the complexity of this algorithm is $O(MN^{J+1})$ where M is the dimension of each observation and N is the number of the observations. We usually choose $J = 2$ resulting in a time complexity of $O(MN^3)$.

For our experiments we use the weighted version of the modified band-depth, Eq. 7, because it results in fewer depth ties compared to the unmodified band-depth. Note that as we are dealing with a generalization of the median, continuity with respect to the regression variable (here, age) *cannot be guaranteed*. Assuming that the underlying data is continuous, a "more continuous" behavior may be achieved using more and



(a) Observations (b) Age histogram

Figure 2: (a) 20 observations generated based on Eq. (12) and colored by age. (b) The age histogram of the observations.

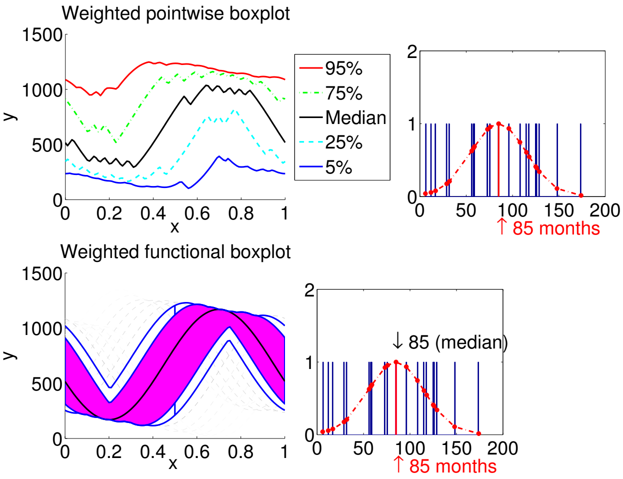


Figure 3: Comparisons of the atlases built by the weighted point-wise boxplot (top) and the weighted functional boxplot (bottom) on the synthetic data. The atlases are adapted to the age of 85 months. The median computed by the weighted point-wise boxplot is a point-wise median, and the median computed by the weighted functional boxplots corresponds to an existing observation at 85 months.

sufficiently dense sampled data. In particular, we would expect that additional samples in sparsely sampled regions of a dataset would result in solutions with less severe discontinuities.

4. Comparisons of boxplots for analysis

4.1. Synthetic data

We compare the atlases built by weighted functional boxplots and those built by 1) weighted point-wise boxplots and 2) functional boxplots, using synthetic observations defined by

$$y_i(x) = 500 * (1 + sin(2\pi x + 0.1\pi i)) + 2 * age_i, \tag{12}$$

where $x \in [0, 1]$, i ranges from 1 to 20 (i.e., we generate 20 observations for analysis, shown in Fig. 2(a)), and age_i is the simulated age corresponding to the ith spike in Fig. 2(b). The age varies from 0 to 200 months, that is, $b_l = 0$ and $b_h = 200$.

4.2. Comparison with weighted point-wise boxplots

The top image of Fig. 3 shows an atlas built with the weighted point-wise boxplot including four typical percentiles and the point-wise median. The weights are computed based on the Gaussian function in Section 2.1 with $\sigma = 30$ months. While the median curve follows the overall population trend, it is not "close" to any of the observations because weighted boxplots, applied in a point-wise manner to a population of functions, disregard the spatial aspect of the functional data. In contrast, our method shown in the bottom image of Fig. 3 1) provides a median curve which corresponds to a curve in the data set, i.e., the one at the age of 85 months, and 2) allows for the computation of *functional* outliers (gray dashed lines) which results in a more robust statistical description for the atlas.

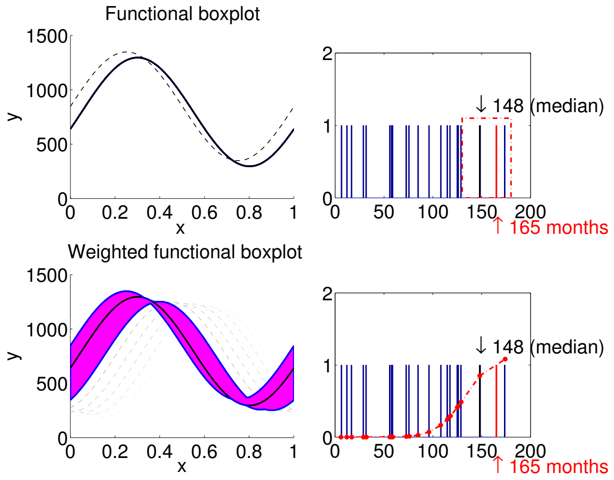


Figure 4: Comparisons of atlases built by the functional boxplot (top) and the weighted functional boxplot (bottom) on the synthetic data. The atlases are built at age 165 months and for both methods the observation at 148 months is selected as the median curve.

4.3. Comparison with functional boxplots

To construct an atlas at a particular age using standard functional boxplots, we use a uniform window to pick curves centered around the age of interest. As shown in Fig. 4, only two curves are available in the uniform window for atlas-building with functional boxplots, and one of them is flagged as an outlier. This atlas includes little information about the population. However, the atlas built using the weighted functional boxplot (with a Gaussian window size comparable to that of the uniform window used in the standard functional boxplots according to Marron and Nolan (1988)) captures the population data much better as it suffers less from the local data sparsity.

For further illustration, we build a set of atlases from the synthetic curves at the age associated to each curve using functional boxplots and weighted functional boxplots. Each age-matched atlas has a median curve, and ideally the age of the atlas matches with the age of the median when the age of the population is evenly distributed, indicated by the cyan dots in Fig. 5. For this synthetic dataset, we want to determine which one provides a better approximation of the median age to the atlas age, the functional boxplot or the weighted functional boxplot. As shown in Fig. 5, the magenta line estimated by the weighted functional boxplot (WFB) is much smoother than the blue line estimated by the functional boxplot (FB), and the magenta line is closest to most of the cyan dots, indicating that the

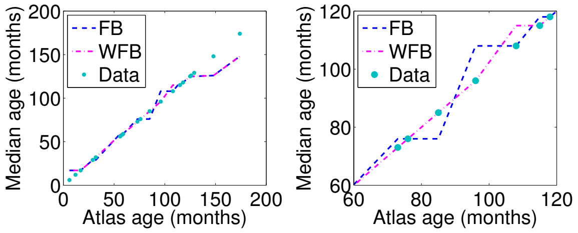


Figure 5: Comparison of the atlas age and the median age between the functional boxplot (FB, the blue dashed line) and the weighted functional boxplot (WFB, the magenta dashed line). The cyan dots show the ideal case, that is, a method has a better performance if it passes through more cyan dots. The right image is a close-up view of the left one.

Table 1: Comparison of the median ages estimated by functional boxplots (FB) and weighted functional boxplots (WFB) on synthetic data.

	FB	WFB
Mean of relative errors	15.25%	**13.99%**
Equal to atlas ages	35%	**50%**
Closer to atlas ages*	5%	**20%**

*This measure counts the frequency with which the estimated median ages are closer to the true age for functional and weighted functional boxplots respectively. In 75% of the cases the median ages from these two methods are identical.

weighted functional boxplot has a better performance than the functional boxplot for spatio-temporal atlas construction.

Table 1 provides quantitative measurements on the median ages computed by the functional boxplot and weighted functional boxplots with respect to the atlas ages. In particular, we evaluate the relative age error, how frequently the methods return a median curve of exactly the correct age and with what frequency the estimated median curve's age is closer to the real age for the functional versus the weighted functional boxplot. The weighted functional boxplot outperforms the standard functional boxplot for all these measures.

5. Comparison with the point distribution model

The point distribution model (PDM) (Cootes et al., 2004) is a powerful method to statistically describe shape variations. Shape variation is captured by computing a mean shape and the principal shape variations around this mean through principal component analysis. It is important to note that the objective of a PDM is different from that of the functional boxplot. Whereas PDM is used to capture the major modes of shape variation through a multi-variate Gaussian distribution, the functional boxplot is free of distributional assumptions as it is a form of order statistics. The functional boxplot is therefore robust to outliers and readily allows for the computation of α-central regions, such as the interquartile range, to quantify data variation.

To demonstrate the difference in behavior between the PDM and the functional boxplot Fig. 6 shows the shape variation of 18 2D hand outlines from Cootes et al. (1995) as captured through a PDM and the functional boxplot. The PDM readily allows for the visualization of principal modes of shape variation, whereas the functional boxplot provides an intuitive way of looking at the spatial differences observable within for example the 50% confidence region (see Section 6.2 for details on how to compute the confidence region). Hence, both methods provide useful, but complementary information.

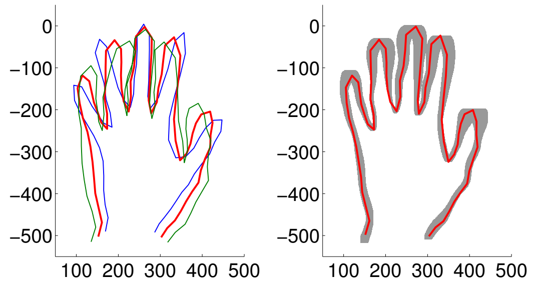


Figure 6: Comparison between the point distribution model (left) and the functional boxplot (right) applied to 18 2D hand contours. Left: mean shape in red and shape variation along the first mode for −3 standard deviations (blue) and for +3 standard deviations (green). Right: median shape in red and 50% confidence region in gray.

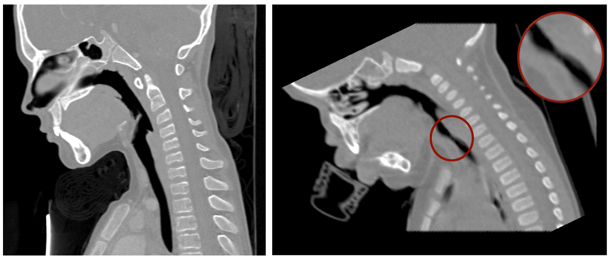


Figure 7: CT scans for a control subject (left, CRL04) and a subglottic stenosis patient (right, SGS03). The zoomed-in part in the red circle shows the location of subglottic stenosis, the narrowing of the airway.

6. Applications

6.1. Real data

In this section we show example applications using the weighted functional boxplot. The examples involve functions, shapes, and images.

Functions. Our first application is the construction of a pediatric airway atlas from normal subjects to assess airway obstruction (i.e. subglottic stenosis, SGS) (Daniel, 2006), as shown in the computed tomography (CT) images in Fig. 7. The observations are a population of 1D functions describing airway cross-sectional areas parameterized along the centerline of the airway as shown in Fig. 9. These functions are generated from 3D CT data for 68 normal subjects using a simplified airway model (Hong et al., 2013b), shown in Fig. 8, followed by a landmark-based spatial alignment (Ramsay and Silverman, 2005). The spatial alignment is based on five key anatomic landmarks: nasal spine, choana, epiglottis tip, true vocal cord and tracheal carina. For each landmark, there is a physical position on the centerline and a mean position of that landmark for all subjects. We estimate a warping function parameterized as a spline smoothly passing through pairs of physical and mean

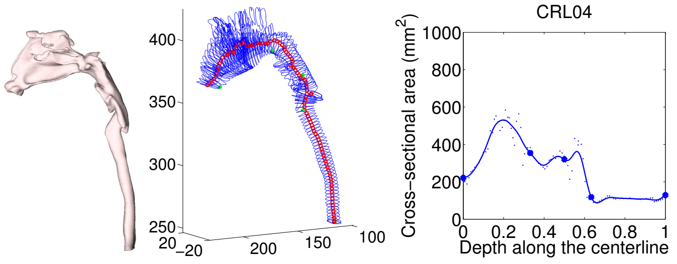


Figure 8: The simplified airway model for converting a 3D airway geometry to a 1D curve. Left: the geometry segmented from a CT image, CRL04; middle: the centerline of the airway with cross sections along the centerline; right: the curve of the cross-sectional area with the depth along the centerline.

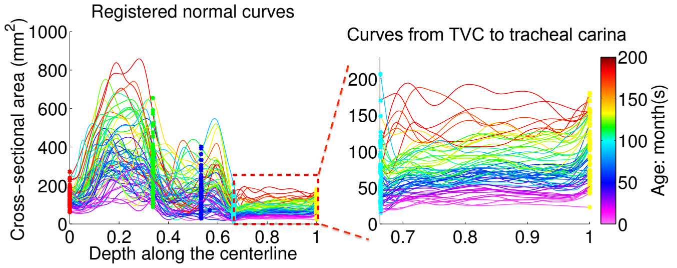


Figure 9: Normal curves for pediatric airway atlas construction, which are registered based on the following five landmarks: nasal spine, choana, epiglottis tip, true vocal cord (TVC) and tracheal carina (from left to right). Zoomed-in: the sub-region from TVC to tracheal carina where the subglottis is located.

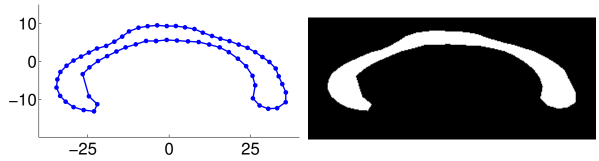
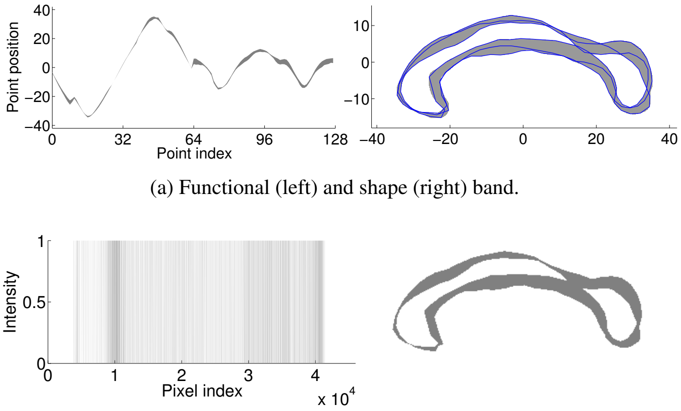


Figure 10: Examples of the corpus callosum shape (left) and the binary image of the corresponding segmentation (right).



(a) Functional (left) and shape (right) band.

(b) Functional (left) and image (right) band.

Figure 11: The functional bands, delimited by three corpus callosum shapes (the blue contours), and their corresponding shape band and image band.

position for all landmarks for the registration of the functions. We focus the analysis on the region between the true vocal cord (TVC) and the tracheal carina where the subglottis is located, as shown in the right image of Fig. 9. The 68 normal functions are used to build a normal control pediatric airway atlas to assess 19 SGS subjects pre- or/and post-surgery.

Shapes. The second application is to build a corpus callosum atlas and to explore its shape changes with age. The observations are a collection of 32 corpus callosum shapes of varying ages from Fletcher (2011). Each shape is represented by 64 2D boundary points as shown in the left image of Fig. 10. We perform affine alignment before atlas constructions.

Images. The third application is to understand age-related changes of the corpus callosum using binary images of the corpus callosum segmentations. The images are converted from the aligned corpus callosum shapes, and one example is shown in the right image of Fig. 10.

6.2. *Functional representation of shapes and images*

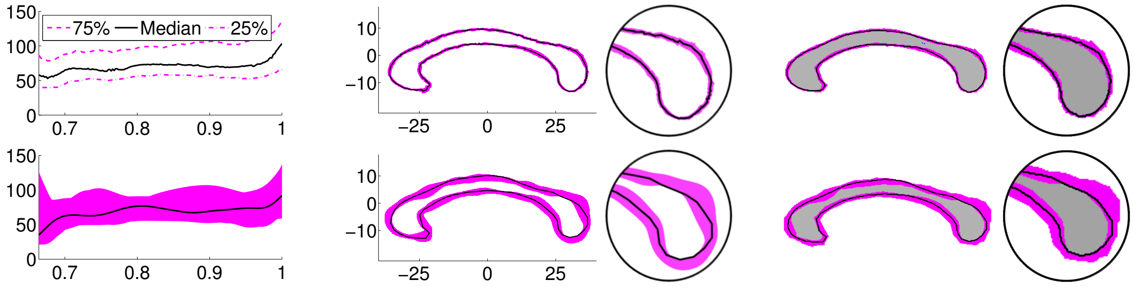
In our experiments, we treat shapes and images as functions by vectorizing the data. After analysis, we convert the functional form back to the original representation of the data objects. It is instructive to look at the effect of this vectorizing step in the context of binary images which we use as an image-based method to represent shapes (contours). In this case, images represent shapes through indicator functions. Note that we discuss curves in 2D in this paper, but the principle extends to the representation of any closed co-dimension one object, e.g., closed surfaces in 3D. Assume the shapes are represented by images through indicator functions $\mathbb{1}_{I_i}$, where I_i is the set which indicates the interior of the shape S_i, i.e., $I_i = \{x : x \text{ inside } S_i\}$ and $\mathbb{1}_{I_i}(x) := 1$ if $x \in I_i$ and 0 otherwise. We can then write intersections and unions of sets through the indicator functions as

$$\mathbb{1}_{S_i \cap S_j} = min\{\mathbb{1}_{S_i}, \mathbb{1}_{S_j}\} \quad \text{and} \quad \mathbb{1}_{S_i \cup S_j} = max\{\mathbb{1}_{S_i}, \mathbb{1}_{S_j}\}. \quad (13)$$

The band-depth defined in Section 2.2.1 is based on evaluating $I\{G(y) \subseteq B(y_1, \cdots, y_j)\}$. Applied to indicator functions this expression is equivalent to

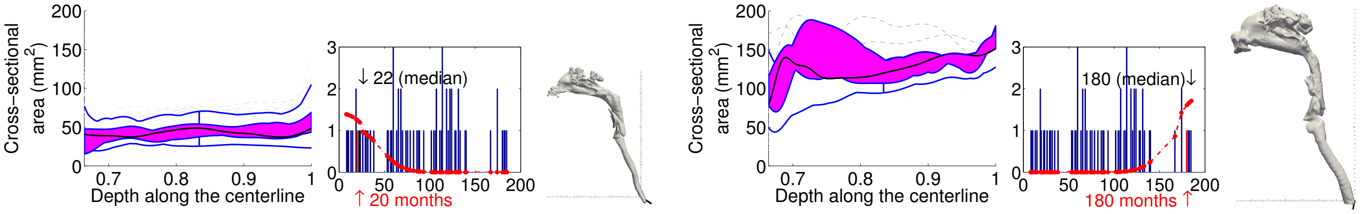
$$I\left\{\bigcap_i S_{y_i} \subseteq S_y \subseteq \bigcup_i S_{y_i}\right\} \quad (14)$$

as the band B is constructed by taking the minima and maxima over all functions. For the indicator functions the minimum and maximum operators correspond to the set intersections and unions respectively (due to the associativity of set union and intersection). Applying the functional boxplot to vectorized indicator functions of images representing shapes (Hong et al., 2013a) is therefore equivalent to the definition of contour boxplots proposed independently by Whitaker et al. (2013), unifying the two methods. Whitaker et al. (Whitaker et al., 2013) introduces contour boxplots to quantify uncertainty in feature sets from simulation ensembles such as for example obtained from fluid simulations. Our weighted functional boxplot can therefore also be interpreted as an extension of the method of Whitaker et al. (Whitaker et al., 2013) to weighted contour boxplots. This shows that the vectorization approach is quite natural in the context of indicator-function-based shape representa-



(a) Functions, pediatric upper airways (b) Shapes, corpus callosum boundaries (c) Images, corpus callosum segmentations

Figure 12: Comparison between point-wise (top) and functional (bottom) boxplots on functions, shapes and images. The black curve is the median and for the point-wise boxplots it is the point-wise median. The magenta region is the 50% confidence region.



(a) Atlas at 20 months and the median airway geometry (b) Atlas at 180 months and the median airway geometry

Figure 13: Age-adapted atlases for functions: pediatric airway atlases at 20 and 180 months respectively. The two airway geometries correspond to the median subjects selected by the age-matched atlases. The older atlas has a larger airway size compared to the younger atlas, indicating the importance of building age-matched atlases.

tion. The relation for the contour-based representations is theoretically less clear. However, our experiments indicate that in practice this method achieves similar results to the indicator-function-based shape representation while being computationally more efficient as the shape representation is more compact.

Specifically, we compute bands for shapes and images as follows:

Shape band. To compute the band for aligned shapes, taking the three blue curves in Fig. 11(a) as an example, we first vectorize them to compute the functional band, shown on the left in Fig. 11(a). Then, for a 2D point on the shape, (p, q), its variation is within the rectangular region with the diagonal given by the two points on the band's boundary, $(min(p), min(q))$ and $(max(p), max(q))$. For a 3D point, its variation is within a rectangular solid. The union of these rectangular regions then forms the shape band. With a sufficiently dense sampling of the functions, we obtain a smooth shape band as illustrated on the right of Fig. 11(a). This shape band contains all three curves.

Image band. Compared with the shape band, the image band is much easier to construct. As discussed above for binary images, the standard functional boxplot theory can be directly applied. Converting the obtained bands back to the image domain immediately results in the desired image band. The image band can in the same way be constructed for non-binary images. Fig. 11 shows the functional band for three binary images of corpora callosa on the left, and the corresponding image band on the right.

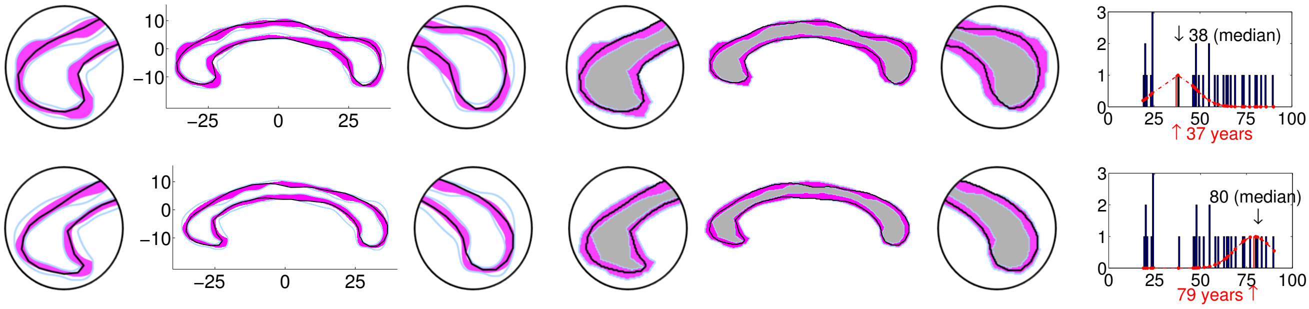
Fig. 11 also shows that both shape and image bands are similar and correctly capture the range of the observations.

6.3. Comparison with point-wise boxplots

We compare the functional boxplot to the point-wise approach on real datasets to demonstrate the advantages of our method. Fig. 12 shows the median (the black curve) and the confidence region (the 50% central region, magenta) for both point-wise and functional boxplots. We count the number of data objects inside the confidence region shown in Table 2: for the point-wise boxplots only 12 (of 68) functions and none of the shapes or images are fully within the confidence region. However, the functional boxplot, by construction, provides a confidence region containing 50% of the data objects. We consider this a more intuitive representation of true data-object variation. To construct the point-wise confidence regions for shapes we locally compute distances with respect to the median point which establishes an (unsigned) ordering. The confidence region is then the convex hull of the closest half of the points. This strategy would extend to constructing approximate confidence regions with respect to manifold embedding coordinates. Specifically, in the coordinate system after manifold embedding, each observation is represented as a point and the median is defined as the point with the minimal sum of the squared geodesic distances to other points on the manifold. The confidence region can then be defined as the convex hull on the manifold formed by half of the points with the closest geodesic distances to the median point. This is conceptually similar to the way we construct shape bands.

6.4. Atlas Construction with weighted functional boxplots

The weighted functional boxplot is used to build a pediatric airway atlas with variance $\sigma = 24$ months for the weighting function, and the corpus callosum shape/image atlases with $\sigma =$



(a) Shapes of corpus callosum

(b) Images of corpus callosum

Figure 14: Age-adapted atlases for shapes and images: corpus callosum atlases at 37 (top) and 79 (bottom) years respectively. Zoomed-in: the anterior (the splenium, on the right of the atlas) and posterior (the genu, on the left of the atlas) portions of corpus callosum atlases. The atlases at different ages, especially the zoomed-in parts, clearly show the thinning of the corpus callosum with age.

Table 2: The number of data objects inside the 50% central region for functions, shapes and images in Fig. 12.

	Functions	**Shapes**	**Images**
Point-wise boxplots	12/68	0/32	0/32
Functional boxplots	**34/68**	**16/32**	**16/32**

The first number is the sum of the data objects inside the central region, and the second number is the total number of observations.

Table 3: Comparison of the median ages estimated from the functional boxplot (FB) and the weighted functional boxplot (WFB) on the pediatric airway dataset.

	FB	**WFB**
Mean of relative errors	19.75%	**15.05%**
Equal to atlas ages	**11.76%**	7.35%
Closer to atlas ages*	20.59%	**26.47%**

*This counts the number of the median ages that are closer to the atlas ages between functional boxplots and weighted functional boxplots; 52.94% of the median ages from these two methods are equal.

10 years. For the pediatric airway application, the age range varies from 0 to 200 months, that is, $b_l = 0$ and $b_h = 200$. For the corpus callosum application, we set the age within (0, 100 years), that is, $b_l = 0$ and $b_h = 100$.

6.4.1. *Pediatric airway atlas*

Fig. 13 shows two pediatric airway atlases at different ages, 20 months and 180 months respectively. The pediatric airway atlases capture increases in cross-sectional airway area with age which is consistent with the growth pattern for pediatric airways and indicates the necessity of building an age-adapted atlas as a reference. Furthermore, in Table 3 we measure the difference of median ages estimated by functional boxplots and weighted functional boxplots. Similar to Section 4.3, we build an age-matched atlas for each control subject and use the age of the selected median subject for comparison. The weighted functional boxplot leads to a smaller mean relative error and more median ages are closer to the atlas ages. However, fewer median ages agree exactly with the atlas ages. This is acceptable because the cross-sectional area of pediatric airways increases with age in general while small variations may be caused, e.g., by measurement errors and difference in true developmental age. Overall, the weighted functional boxplot performs well at building the pediatric airway atlas.

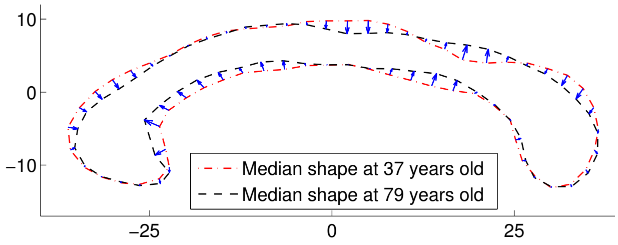


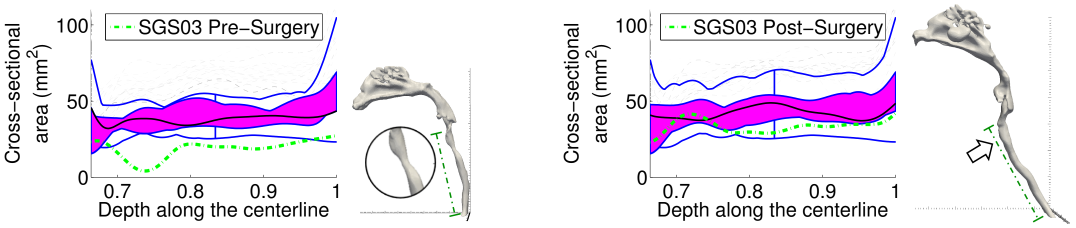
Figure 15: The median shapes of two corpus callosum atlases at different ages and the direction of change of the corresponding points on the boundaries.

6.4.2. *Corpus callosum atlas*

In Fig. 14, we select atlases at age 37 and 79 years for both the shape and the segmentation of corpus callosum to demonstrate atlas changes with respect to age. The two corpus callosum atlases reveal the thinning in the shape and the decreasing volume in the image with age, especially at the anterior (the splenium) and posterior (the genu) portions consistent with (Driesen and Raz, 1995; Fletcher, 2011). To further visualize these changes, we overlap the median shapes of the corpus callosum atlases in Fig. 14, and display the directions of change for all corresponding points on the boundary in Fig. 15. Most parts of the median shape, especially the anterior and posterior regions, show the thinning of the corpus callosum with age.

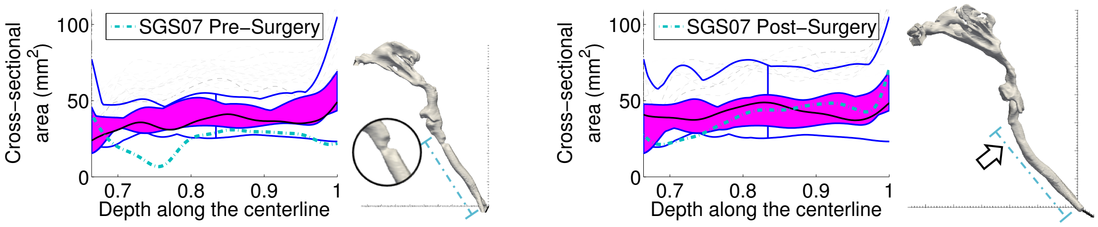
7. Computational cost for building a statistical atlas

The algorithm is implemented in Matlab. All the experiments were run on an Intel® Xeon(R) CPU E5645 system with 2.4GHz. Table 4 shows the computation times for building atlases from populations of observations with different numbers of datasets and different numbers of 1D/2D points and pixels. Most experiments required less than one second of runtime. The image-based approach, while still reasonably fast, is as expected the slowest as the dataset size is largest.



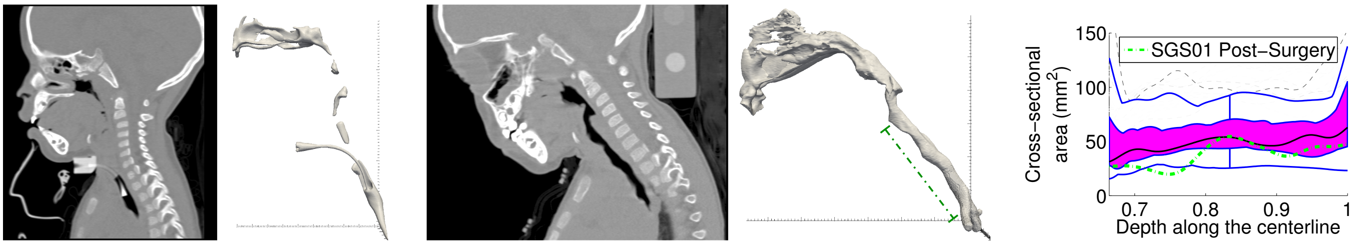
(a) Pre-surgery: the age-matched atlas and the airway geometry (b) Post-surgery: the age-matched atlas and the airway geometry

Figure 16: Airway changes for a subject, SGS03, pre- and post-surgery (green dashed lines) compared to the age-matched atlas. The stenosis of the airway is marked by the zoomed-in circle on the pre-surgery geometry and no visible stenosis exists in the post-surgery geometry (the arrow on the right image corresponding to the subglottic area).



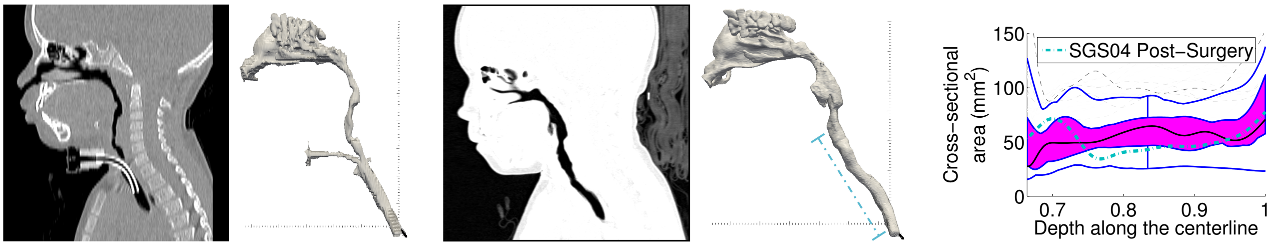
(a) Pre-surgery: the age-matched atlas and the airway geometry (b) Post-surgery: the age-matched atlas and the airway geometry

Figure 17: Airway changes for a subject, SGS07, pre- and post-surgery (cyan dashed lines) compared to the age-matched atlas. The stenosis of the airway is marked by the zoomed-in circle on the pre-surgery geometry and no visible stenosis exists in the post-surgery geometry (the arrow on the right image corresponding to the subglottic area).



(a) Pre-surgery CT scan and airway geometry (b) Post-surgery CT scan and airway geometry (c) Age-matched atlas for post-surgery

Figure 18: Airway changes for SGS01 pre- and post-surgery. Before surgery there is a tracheostomy tube in the airway. After surgery the subglottic stenosis is resolved. Compared with the age-matched atlas most of the corresponding curve is within the maximal non-outlying envelope, indicating a successful surgery.



(a) Pre-surgery CT scan and airway geometry (b) Post-surgery CT scan and airway geometry (c) Age-matched atlas for post-surgery

Figure 19: Airway changes for SGS04 pre- and post-surgery. Before surgery there is a tracheostomy tube in the airway. After surgery the subglottic stenosis is resolved. Compared with the age-matched atlas all of the corresponding curve is within the maximal non-outlying envelope, indicating a successful surgery.

Table 4: The computational cost of building an atlas based on the weighted functional boxplot.

	Observations		Computational
	Size	Number	Cost (s)
Synthetic functions	101 1D points	20	0.024
Pediatric airway	169 1D points	68	0.50
Corpus callosum (shape)	64 2D points	32	0.34
Corpus callosum (image)	157 × 456 pixels	32	29.56

8. Assessment with statistical atlas

8.1. Comparison pre- and post-surgery

To test the utility of the statistical atlas built by weighted functional boxplots we show the airway changes of two SGS subjects before and after surgery compared to the age-matched normal control airway atlases. The subject shown in Fig. 16 is SGS03, a male who had two CT scans, one before surgery at 9 months and the other after surgery at 20 months. Fig. 17 shows another male (SGS07) who had a CT scan before surgery at 6 months and another one after surgery at 15 months. Before treatment, there was a constricted region outside the atlas for both children, corresponding to the dip in the cross-sectional area curve and the zoomed-in circle of the geometry in both Fig. 16(a) and Fig. 17(a). After treatment, the airway size increased and the corresponding curve is almost entirely within the maximal non-outlying envelope of the atlas. Also there is no visible stenosis in the geometry as shown in both Fig. 16(b) and Fig. 17(b), indicating that the surgeries for these children were successful.

Fig. 18 and Fig. 19 show two additional children with subglottic stenosis. We can see the tracheostomy tubes in the CT scans and the airway geometries. We do not compute the cross-sectional areas for such cases, because their airways appear disconnected before surgery and breathing is accomplished through the tracheostomy tubes. Minimal cross-sectional areas are set to zero. After surgery, the airway cross-sectional areas greatly increase. For SGS01 only a small part of the corresponding curve is slightly outside the maximal non-outlying envelope of the atlas, and for SGS04 its whole corresponding curve is totally inside the maximal non-outlying envelope, indicating successful surgeries also for these two cases.

8.2. Quantitative measurements

8.2.1. Definition of the scoring system

The Myer-Cotton grading system (Myer et al., 1994) is commonly used in clinical diagnosis for estimating the severity of subglottic stenosis in the pediatric upper airway. It describes the stenosis by the relative percentage reduction of the cross-sectional area at the subglottis. In practice, this is determined by using different sizes of endotracheal tubes. Similar to the Myer-Cotton system, we define a scoring system based on the age-matched atlas to quantitatively measure the severity of subglottic stenosis for the pediatric upper airway. Compared with the Myer-Cotton system, our measurement is non-invasive and not limited by the size of the endotracheal tubes.

For each individual curve y, from TVC to tracheal carina, we build an atlas that is adapted to the age of the corresponding subject, and compare it with the minimal curve of the atlas, C_{lower_fence}, because this minimal curve can be considered the minimal cross-sectional area of a *normal* airway at that age. With the minimal curve as the reference of the airway's cross-sectional area, our scoring system is defined as:

$$Score(\mathrm{y}) = \min_x((\mathrm{y}(x) - C_{lower\_fence}(x))/C_{lower\_fence}(x)). \quad (15)$$

If the whole curve y is above the minimal curve, the score will be non-negative; otherwise it will be negative. Since all $\mathrm{y}(x) \geq 0$, the lower bound for the score is -1. While there is no theoretical upper bound to this definition, the score will be upper-bounded in practice by the largest observable cross-sectional areas for a given age. That is, the score of the cross-sectional area curve for a pediatric upper airway is within $[-1, \infty)$, where -1 indicates a fully closed airway with a zero cross-sectional area somewhere. A negative score indicates a potential stenosis and a normal control subject usually has a non-negative score. Overall, the higher the score, the more normal the corresponding subject will be. Note that our measurement is not directly comparable to the Myer-Cotton system, as the Myer-Cotton system computes within-subject scores by estimating the cross-sectional area of what should be considered a non-constricted airway. Our score on the other hand makes use of population data contained in the normal control atlas to define what a minimum normal cross-sectional area should be. Nevertheless, the two scoring systems can be made "roughly comparable" by setting all positive atlas-derived scores to zero (indicating a healthy airway) and negating all negative scores.

8.2.2. Scores for all subjects

Based on our scoring system, we score the pediatric upper airways not only for SGS patients but also for the normal controls. The scores shown in Fig. 20 are estimated based on the atlases built by weighted point-wise boxplots, functional boxplots, and weighted functional boxplots. The subjects shown in the plots include 68 normal controls and 17 SGS patients (6 pre-surgery, 11 post-surgery). Among the total 19 SGS patients two subjects that have completely obstructed airways are directly scored as -1 and are not shown in Fig. 20. Within the 11 post-surgery subjects, some of them have no stenoses after surgery, others show improvement in the airway but still exhibit slight stenoses.

To verify whether there is a statistically significant score difference among groups, we use two types of hypothesis tests, the two sample t-test (Snedecor and Cochran, 1989) with the normal distribution assumption for samples, and the Wilcoxon rank sum test (Siegel, 1956), a non-parametric statistical hypothesis test for populations that cannot be assumed to be normally distributed. Table 5 shows the testing results among the following three groups: SGS pre-surgery, SGS post-surgery, and control subjects. We use three different analysis approaches: weighted point-wise boxplots, functional boxplots, and weighted functional boxplots. In each test between two groups the smallest

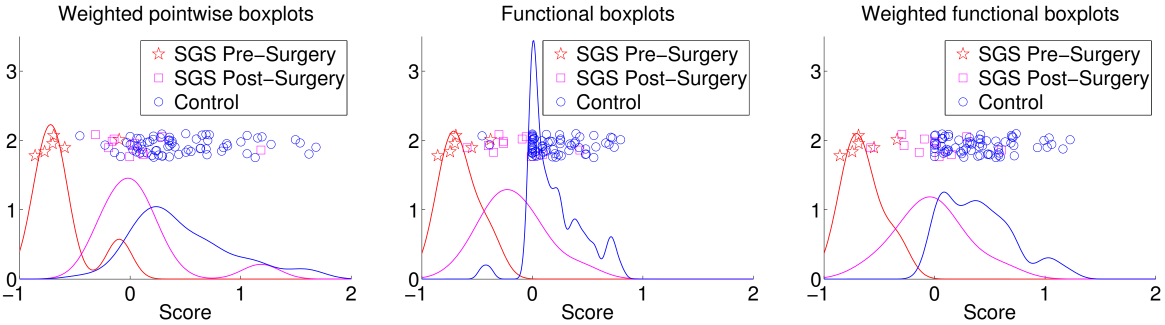


Figure 20: The scores for all subjects, including three groups, SGS pre-surgery, SGS post-surgery and control subjects, based on the atlases built by weighted pointwise boxplots, functional boxplots, and weighted functional boxplots (from left to right). The curves in different colors represent the kernel density estimations for different groups. Note: the y-axis in the plots is a random height to visualize the scores clearly.

Table 5: P-values of two types of tests on the scores for pediatric upper airways estimated based on weighted point-wise boxplots, functional boxplots and weighted functional boxplots.

	Two sample t-test			**Wilcoxon rank sum test**		
	Weighted point-wise boxplots	Functional boxplots	Weighted functional boxplots	Weighted point-wise boxplots	Functional boxplots	Weighted functional boxplots
Pre v.s. CRL	2.1e-07	**1.4e-11**	2.6e-11	7.2e-05	6.0e-05	**5.6e-05**
Pre v.s. Post	1.7e-03	1.4e-03	**5.2e-04**	1.9e-03	1.1e-03	**6.5e-04**
Pre v.s. Post&CRL	7.7e-07	1.7e-09	**1.4e-09**	8.2e-05	6.7e-05	**5.7e-05**
Post v.s. CRL	9.6e-03	**4.4e-05**	9.7e-05	1.6e-03	**9.9e-05**	3.9e-04

Notes: Pre represents the SGS pre-surgery group, Post represents the SGS post-surgery group, CRL represents the normal control group, and Post&CRL represents the union of the SGS post-surgery and normal control groups.

p-value is in highlighted boldface in Table 5. Overall, the results suggest that the weighted functional boxplot is superior to the standard functional boxplot and the weighted pointwise-boxplot in separating the SGS pre-surgery subjects from the SGS post-surgery subjects or/and the normal controls, though all results are highly statistically significant. Note that it is not obvious that post-surgery and normal control subjects should be well distinguishable as a successful surgery should result in a post-surgery airway which should be close to normal.

A closer look at the scores resulting from the three different analysis methods reveals that the scores for the weighted point-wise boxplot (shown in Fig. 20(left)) mix the SGS post-surgery subjects with the normal controls. While this could be desired, as a successful surgery should result in a more "normal-looking" airway, more importantly the weighted point-wise boxplot assigns negative weights to some of the normal controls. This suggests potential stenoses in the control airways and conflicts with the definition of our scoring system. This negative score effect for normal controls is not present for the weighted functional analysis approach, but also appears when using the un-weighted functional analysis (see details below). This suggests that the weighted functional analysis is more appropriate for this application.

Considering the scores of the functional boxplot (Fig. 20 (middle)), there are two normal control subjects scored with negative values: CRL32 and CRL102, whose curves and age-matched atlases are shown in Fig. 21. For these two subjects, parts of their curves are below the atlases built by the functional boxplot. The age of CRL32 is 8 months which is near the lower age boundary with limited information for atlas-building, and CRL102 is at the age of 182 months and therefore suffering from local data sparsity in our current dataset. Compared with the functional boxplot, the weighted functional boxplot shows better performance given the limited data information and the local data sparsity as also shown in Section 4. The curves for these two subjects are fully within the atlases built by weighted functional boxplots and scored with non-negative values, as shown in the right column of Fig. 21.

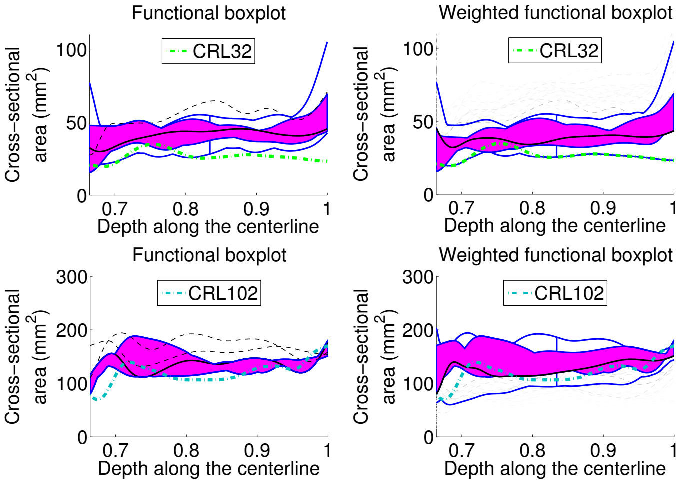


Figure 21: Two control subjects, represented by colored dashed curves and their age-matched atlases. The curves obtain negative scores when using the functional boxplot and non-negative scores when using the weighted functional boxplot.

Table 6: Comparison of the scores for SGS subjects using three different methods with the clinical diagnosis based on the Myer-Cotton grading system.

Patient Id	Surgery	Myer-Cotton	WPB	FB	WFB
SGS03	Pre	80-90%	86.0%	85.6%	85.6%
SGS07	Pre	85%	77.4%	74.5%	74.5%
SGS11	Pre	50%	59.2%	54.8%	54.8%
SGS12	Pre	70%	70%	70%	68.7%
SGS13	Pre	60-70%	**9.9%**	**37.8%**	**34.0%**
SGS18	Pre	60-70%	69.2%	69.2%	68.8%
SGS03_V3	Post	0%	**0.3%**	**1.1%**	0%
SGS05	Post	0%	0%	0%	0%
SGS06	Post	40-50%	**0%**	35.1%	**13.9%**
SGS08	Post	0%	0%	0%	0%
SGS09	Post	50%	**19.6%**	59.1%	57.8%
SGS10	Post	grade I	**0%**	39.9%	27.6%
SGS14	Post	50%	**0%**	**26.1%**	**0%**
SGS17	Post	0%	**16.3%**	0%	0%
SGS07_V3	Post	30%	**14.5%**	**9.3%**	**9.3%**
SGS04_V3	Post	grade I: 10%	**0%**	5.6%	**0%**
SGS01_V3	Post	15-20%	31.4%	30.2%	29.6%

Notes: Weighted point-wise boxplots (WPB), functional boxplots (FB), weighted functional boxplots (WFB). The scores are converted based on the correspondence between our scoring system and the Myer-Cotton system in Section 8.2.1. Grade I represents an obstruction within (0% - 50%].

8.2.3. *Score comparison of pre- and/or post-surgery*

Table 6 shows the scores for SGS subjects using weighted point-wise boxplots, functional boxplots, and weighted functional boxplots, and compares them with the clinical diagnosis based on the Myer-Cotton grading system. The scores computed using our scoring system are converted to be roughly comparable to the corresponding Myer-Cotton values as described in Section 8.2.1. Scores that are outside of the ±20% deviation of the clinical diagnosis and that are zero for subjects with stenoses or non-zero for subjects without stenoses are shown in boldface. Table 6 shows that the weighted point-wise boxplot frequently gives results which are not what would be expected from the Myer-Cotton scores. The scores based on the functional boxplot and the weighted functional boxplot both give results which are more comparable to the Myer-Cotton scoring.

To further reveal the differences between functional boxplots and weighted functional boxplots, we quantitatively compare the four SGS subjects pre- and post-surgery shown before in Section 8.1. In general, as shown in Fig. 22 the scores of these four subjects increase after surgery for both methods, which indicates that all subjects' airways improved from the surgeries.

8.2.4. *Classification of Control and SGS subjects*

To demonstrate the classification accuracy for separating SGS pre-surgery subjects from SGS post-surgery and/or control subjects we compute the confusion matrices (Provost and Kohavi, 1998) based on the scores estimated from weighted point-wise boxplots, functional boxplots and weighted functional boxplots, as shown in Table 7. We label the SGS pre-surgery subjects as positive (P), and both SGS post-surgery and control subjects as negative (N). We repeatedly take one subject out for testing (i.e., leave-one-patient-out) and trained a Support Vector Machine (SVM) (Cortes and Vapnik, 1995) on the data from the remaining subjects. We use a linear SVM for our experiments. For the confusion matrices, we calculate the numbers of true positive (TP), true negative (TN), false positive (FP), false negative (FN) instances. Besides, we use the true positive rate (TPR = TP/(TP+FN)), the false positive rate (FPR = FP/(FP+TN)), the positive predictive value (PPV = TP/(TP+FP)), and the accuracy (ACC = (TP + TN)/(P+N)) to further assess the performance of the classification between SGS pre-surgery subjects and others.

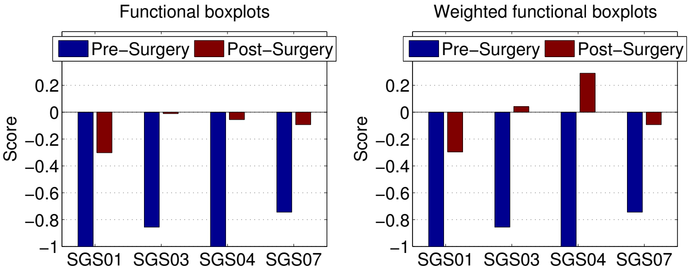


Figure 22: Quantitative comparison of the scores for four SGS subjects before and after surgery using functional boxplots and weighted functional boxplots for atlas-building.

In the classification between SGS pre-surgery and control subjects, for weighted point-wise boxplots one SGS pre-surgery subject is regarded as a normal control and one control subject is regarded as pre-surgery; for functional boxplots there are two false positive subjects, which means two children test positive but actually do not have subglottic stenoses. In contrast, the weighted functional boxplot result shows no false positives or false negatives and yields 100% classification accuracy. In the classification between SGS pre- and post-surgery subjects, the weighted point-wise boxplot has one misclassified subject, and both functional boxplots and weighted functional boxplots have one false positive and one false negative. The misclassified cases will be discussed in detail in the next section. The accuracy of classifying the pre- and post-surgery subjects using the weighted functional boxplot is about 88%. If we combine the SGS post-surgery subjects with the normal controls, the accuracy of the weighted functional boxplot increases to about 96%, which is higher than that of the functional boxplot.

It is important to note that no fully conclusive statements can be made based on the presented classification results. While Table 7 indicates better prediction performance when using WFBs, further tests with larger sample sizes are needed to substantiate our claims.

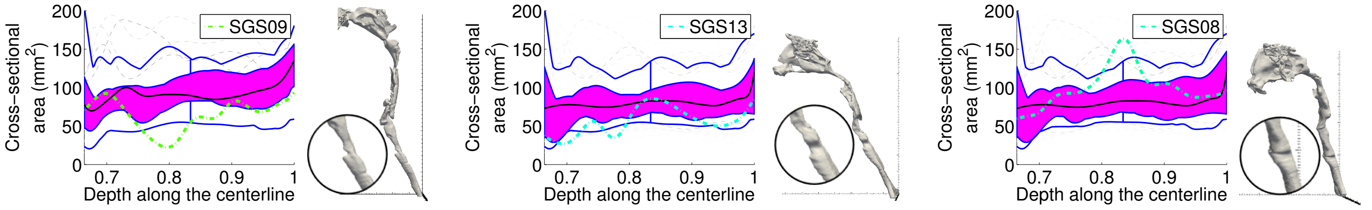
8.2.5. *Discussion of SGS outliers*

Based on the above discussion, the scores computed from weighted functional boxplots can be used to roughly divide the pediatric upper airways into three different groups: the SGS pre-surgery group (score in [-1, -0.5)), the SGS post-surgery group (score in [-0.5, 0)), and the normal control group with a score larger or equal to zero. In this classification, three representative subjects should be discussed. Namely, SGS09, SGS13 and SGS08, which are shown in Fig. 23 together with their cross-sectional area curves in the age-matched atlases and

Table 7: The confusion matrices among groups: SGS pre-surgery (Pre), SGS post-surgery (Post), and control (CRL).

	Pre (P) v.s. CRL (N)		**Pre (P) v.s. Post (N)**		**Pre (P) v.s. Post&CRL (N)**	
Weighted point-wise boxplots	TP = 5	FP = 1	TP = 5	FP = 0	TP = 5	FP = 2
	FN = 1	TN = 67	FN = 1	TN = 11	FN = 1	TN = 77
	TPR = 0.83, FPR = 0.01		TPR = 0.83, FPR = 0.0		TPR = 0.83, FPR = 0.03	
	PPV= 0.83, ACC = 0.97		PPV = 1.0, ACC = 0.94		PPV = 0.71, ACC = 0.96	
Functional boxplots	TP = 6	FP = 2	TP = 5	FP = 1	TP = 5	FP = 8
	FN = 0	TN = 66	FN = 1	TN = 10	FN = 1	TN = 71
	TPR = 1.0, FPR = 0.03		TPR = 0.83, FPR = 0.09		TPR = 0.83, FPR = 0.10	
	PPV= 0.75, ACC = 0.97		PPV = 0.83, ACC = 0.88		PPV = 0.38, ACC = 0.89	
Weighted functional boxplots	TP = 6	FP = 0	TP = 5	FP = 1	TP = 6	FP = 3
	FN = 0	TN = 68	FN = 1	TN = 10	FN = 0	TN = 76
	TPR = 1.0, FPR = 0.0		TPR = 0.83, FPR = 0.09		TPR = 1.0, FPR = 0.04	
	PPV=1.0, ACC = 1.0		PPV = 0.83, ACC = 0.88		PPV = 0.67, ACC = 0.96	

Notes: P (positive), N (negative), TP (true positive), FP (false positive), FN (false negative), TN (true negative), TPR (true positive rate), FPR (false positive rate), PPV (positive predictive value), ACC (accuracy).



(a) Atlas for SGS09 and its airway geometry

(b) Atlas for SGS13 and its airway geometry

(c) Atlas for SGS08 and its airway geometry

Figure 23: Three outliers in Fig. 20 for both functional boxplots and weighted functional boxplots. (a) SGS09 is post-surgery while having a low score more consistent with a pre-surgery subject; (b) SGS13 is pre-surgery while mixed into the post-surgery group; (c) SGS08 is post-surgery appearing as a normal control subject consistent with near normal post-operative airway.

their airway geometries.

SGS09 shown in Fig. 23(a) corresponds to the false positive subject in the confusion matrix of the weighted functional boxplot in Table 7. This subject is post-surgery with a 50% airway obstruction based on the clinical diagnosis. From the cross-sectional area curve and the zoomed-in part of the geometry we can clearly see the subglottic stenosis with a score of -57.8%, thus resulting in being classified as a SGS pre-surgery case, which is sensible.

SGS13 is a pre-surgery subject and according to the clinical diagnosis has a $60 - 70\%$ obstruction in the airway. From Fig. 23(b), we see two stenoses in the airway, as confirmed by the surgeon. However, because of its score, -34.0%, it is the false negative subject in the confusion matrix of the weighted functional boxplot in Table 7 and it is classified as belonging to the SGS post-surgery group.

The last case, SGS08, is post-surgery and has a very high score of 60.5%, indicating a normal subject. As shown in Fig. 23(c), it has a comparable airway size to the atlas and its airway geometry also indicates no stenosis existing in the airway. This subject is confirmed by the surgeon as near normal caliber airway and hence could also be sensibly classified as normal. This case shows that our scoring system can reliably be used to assess abnormalities in pediatric upper airways.

9. Conclusion

We proposed a general method to compute weighted functional boxplots and used it for spatio-temporal atlas building. We applied it to construct a pediatric airway atlas to assess children with subglottic stenosis and a corpus callosum atlas capturing the impact of aging. We also defined a scoring system for pediatric airways based on the statistical atlas to quantitatively measure the severity of subglottic stenosis in children. The proposed method is general, easy to compute, and allows robust statistical description of functional, shape, and image data.

Future work will focus on accounting for the multi-dimensional nature of shapes and images which is currently not considered in our method.

Acknowledgement

This work was supported by NIH 5P41EB002025-28, NIH 1R01HL105241-01, NSF EECS-1148870, NSF EECS-0925875, and NIH 2P41EB002025-26A1. The authors would like to thank Amanda Gee, Mark Hall, and Katherine Abode for their help in this paper. We also thank Tom Fletcher for providing some of the real data used above and all our colleagues for their support in the development of these ideas.

References

Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., Rueckert, D., 2009. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. NeuroImage 46, 726–739.

Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., 1995. Active shape models their training and application. Comp Vision and Image Understanding 61, 3859.

Cootes, T.F., Taylor, C.J., et al., 2004. Statistical models of appearance for computer vision. Imaging Science and Biomedical Engineering, University of Manchester, Manchester M13 9PT, UK March 8.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297.

Daniel, S., 2006. The upper airway: Congenital malformations. Pediatric Respiratory Reviews 7S, S260–S263.

Davis, B.C., Fletcher, P.T., Bullitt, E., Joshi, S., 2010. Population shape regression from random design data. International journal of computer vision 90, 255–266.

Driesen, N., Raz, N., 1995. The influence of sex, age, and handedness on corpus callosum morphology: a meta-analysis, in: Psychobiology, pp. 240–247.

Fletcher, P., Venkatasubramanian, S., Joshi, S., 2009. The geometric median on Riemannian manifolds with application to robust atlas estimation. NeuroImage 45, S143–S152.

Fletcher, T., 2011. Geodesic regression on Riemannian manifolds, in: 3rd MICCAI workshop on mathematical foundations of computational anatomy, pp. 75–86.

Frigge, M., Hoaglin, D.C., Iglewicz, B., 1989. Some implementations of the boxplot. The American Statistician 43, 50–54.

Gerber, S., Tasdizen, T., Fletcher, P.T., Joshi, S., Whitaker, R., 2010. Manifold modeling for brain population analysis. Medical image analysis 14, 643–653.

Hong, Y., Davis, B., Marron, J.S., Kwitt, R., Niethammer, M., 2013a. Weighted functional boxplot with application to statistical atlas construction. Proceedings of the 16th international conference on medical image computing and computer assisted intervention (MICCAI) Part III, LNCS 8151, 584–591.

Hong, Y., Niethammer, M., Andruejol, J., Kimbel, J., Pitkin, E., Superfine, R., Davis, S., Zdanski, C., Davis, B., 2013b. A pediatric airway atlas and its application in subglottic stenosis, in: International symposium on biomedical imaging: from nano to macro, pp. 1194–1197.

Jones, M.C., 1993. Simple boundary correction for kernel density estimation. Statistics and Computing 3, 135–146.

Joshi, S., Davis, B., Jomier, M., 2004. Unbiased diffeomorphic atlas construction for computational anatomy. Neuroimage 23, S151–S160.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th international joint conference on Artificial intelligence 2, 1137–1143.

Liu, R., Parelius, J., Singh, K., 1999. Multivariate analysis by data depth: descriptive statistics, graphics and inference. The annals of statistics 27, 783–858.

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

Marron, J., Nolan, D., 1988. Canonical kernels for density estimation. Statistics and probability letters 7, 195–199.

Marron, J., Ruppert, D., 1994. Transformations to reduce boundary bias in kernel density estimation. Journal of the royal statistical society 56, 653–671.

Myer, C.r., O'Connor, D., Cotton, R., 1994. Proposed grading system for subglottic stenosis based on endotracheal tube sizes. Ann Otol Rhinol Laryngol 103(4 Pt 1), 319–323.

Provost, F., Kohavi, R., 1998. On applied research in machine learning, in: Machine learning, pp. 127–132.

Ramsay, J., Silverman, B., 2005. Functional Data Analysis. Springer.

Schuster, E., 1985. Incorporating support constraints into nonparametric estimators of densities. Communications in Statistics - Theory and Methods 14, 1123–1136.

Siegel, S., 1956. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill.

Snedecor, G.W., Cochran, W.G., 1989. Statistical Methods, Eighth Edition. Iowa State University Press.

Sun, Y., Genton, M., 2011. Functional boxplots. Journal of Computational and Graphical Statistics 20, 316–334.

Wand, M., Jones, M., 1994. Kernel Smoothing. CRC Press.

Wang, H., Marron, J.S., 2007. Object oriented data analysis: Sets of trees. The Annals of Statistics 35, 1849–1873.

Whitaker, R.T., Mirzargar, M., Kirby, R.M., 2013. Contour boxplots: A method for characterizing uncertainty in feature sets from simulation ensembles. Visualization and Computer Graphics, IEEE Transactions on 19, 2713–2722.