Institut für Computergraphik und Algorithmen
Technische Universität Wien
Karlsplatz 13/186/2
A-1040 Wien
AUSTRIA
Tel: +43 (1) 58801-18601
Fax: +43 (1) 58801-18698

Institute of Computer Graphics and Algorithms
Vienna University of Technology
email:
technical-report@cg.tuwien.ac.at
other services:
http://www.cg.tuwien.ac.at/

TECHNICAL REPORT

Statistics-Driven Localization of Dissimilarities in Data

Alexey Karimov, Gabriel Mistelbauer, Thomas Auzinger, and Eduard Gröller

TR-186-2-16-1
April 2016

Statistics-Driven Localization of Dissimilarities in Data

Alexey Karimov, Gabriel Mistelbauer, Thomas Auzinger, and Eduard Gröller

April 30, 2016

Abstract

The identification of dissimilar regions in spatial and temporal data is a fundamental part of data exploration. This process takes place in applications, such as biomedical image processing as well as climatic data analysis. We propose a general solution for this task by employing well-founded statistical tools. From a large set of candidate regions, we derive an empirical distribution of the data and perform statistical hypothesis testing to obtain p-values as measures of dissimilarity. Having p-values, we quantify differences and rank regions on a global scale according to their dissimilarity to user-specified exemplar regions. We demonstrate our approach and its generality with two application scenarios, namely interactive exploration of climatic data and segmentation editing in the medical domain. In both cases our data exploration protocol unifies the interactive data analysis, guiding the user towards regions with the most relevant dissimilarity characteristics. The dissimilarity analysis results are conveyed with a radial tree, which prevents the user from searching exhaustively through all the data.

1 Introduction

A fundamental task in data analysis is the evaluation of internal data consistency by identifying similar and dissimilar subregions inside spatial and temporal data. Manual inspection is generally time-consuming or infeasible due to the amount of data usually encountered in application scenarios such as biomedical image processing, climate studies, *etc.* Thus, the user requires (semi-)automatic search tools. A vast array of domain-specific data analysis tools is available from a large body of research.

In this work we propose a general framework for this task based on sound statistical methods. Thus, our focus lies in the proper and efficient use of statistics in data analysis while still providing sufficient generality of the method with respect to application domains. Having constructed empirical data-value distributions from various regions in the input data, we perform a **statistical hypotheses testing** to compute a **p-value** per region that is used as a dissimilarity measure. We discuss related statistical aspects and provide a practical approach for region aggregation to arrive at a hierarchical ordering of the relative dissimilarities in the data.

We present **visualization means** to assist the user in an interactive exploration of the dissimilarities in the input data. We formulate a **data exploration protocol** which takes advantage of available statistical information on the dissimilarities. A **radial tree** of the hierarchical statistical data guides the user towards potential regions of interest. In case of data change, the global data consistency is tracked via a suitable **timeline plot**.

We apply our framework to two qualitatively different scenarios: temporal data exploration and segmentation editing of volumetric data. We provide the realizations of our abstract concepts in the specific domains. An evaluation of our method's efficacy is given.

2 Related Work

Several works in the visualization and image processing domains involve statistical comparison of regions in volume data in order to detect features or highlight irregularities. In particular, the task of classifying samples of volume data into certain features or materials usually involves such a comparison. Kniss *et al.* [1] discuss a statistical classification procedure that they later apply to rendering. Using a classifier, the method assigns to each sample a probabilistic likelihood that it exhibits certain features of the data. The classifier may have parameters which are evaluated on a training set. The probabilistic likelihoods are then transformed to the posterior probabilities with the Bayes Rule. Based on the probabilities, each sample is fuzzily classified. Each class exhibits various attributes, *e.g.*, transfer functions. This technique provides an efficient classification of the features, but requires classifiers, parameter estimation, feature models, and prior class probabilities beforehand. Our approach

requires only the definition of regions in order to statistically compare them with a set of references.

Tasdizen *et al.* [2] improve tissue classification in MRI (Magnetic Resonance Imaging) by using the data-value distribution in a certain neighborhood around a voxel. They compute the probabilities of observing particular data values in the neighborhood if the voxel belongs to a certain class. The entropy is calculated over the probabilities of the voxel to belong to each possible class. It is iteratively minimized while the classification of voxels changes. This method requires an initial classification that the authors obtain by co-registering the data with a digital atlas. In our approach we do not need this potentially complex step, as we get the references for the statistical comparison directly from the input data.

Lundström *et al.* [3] use partial range histograms to introduce a classification certainty as the second dimension in 2D transfer functions. These histograms collect data values of a certain range in a cubic neighborhood. The ranges of values are determined by fitting Gaussian distributions to the global histogram of the data. The partial range histograms are successively subtracted from the global histogram. While this method improves the classification in regions where ranges of different materials or tissues overlap, it does not explicitly test statistical hypotheses, *i.e.*, whether certain materials or tissues are present in the region of interest. With our statistical approach one can test such hypotheses and make decisions on test results. Heinzl *et al.* [4] extend this work by calculating probabilities of different materials at each data sample. However, a Gaussian distribution of the data values is still assumed.

Johnson and Huang [5] detect features with distribution queries. The data used for forming a query is sampled in the neighborhood of each voxel. Then, the user specifies intervals of data values that are of interest. The filtered data is organized into a histogram that represents the target distribution. The user composes a query with clauses on histogram bins. It is fuzzily matched with the actual volume data. The queries may involve statistics on the histograms, such as standard deviation, skewness, covariance, *etc.* The user assigns to the queries various attributes, *e.g.*, transfer functions. With the rendered image, the user may check hypotheses on the volume data, formulated in terms of the queries. However, this work does not include any statistical tests, which, in certain scenarios, are required to make decisions based on the volume data.

Saad *et al.* [6] investigate the uncertainty of a segmentation compared to expert segmentations in medical image processing. They first construct an atlas that involves two kinds of histograms: likelihood versus shape and likelihood versus appearance. The atlas is constructed with training data from experts. A multivariate Gaussian distribution then models the appearance of the features in the volume data. Prior to the uncertainty analysis, the input-volume data is registered to the atlas. Using the Bayes theorem, voxels are classified with the first-best-guess logic. Finally, the discrepancies between actual segmentation and atlas data are conveyed to the user. This reveals regions of misclassification and abnormalities in the volume data. We propose to capture shape information via a special definition of regions that will be explained in Section 5.2. We collect the data this way instead of using prior shape information stored in an atlas and which is only applicable after co-registration.

Haidacher *et al.* [7] assume that the data values have Gaussian distributions, whose parameters they iteratively evaluate while growing a sphere around each voxel. They use the Jarque-Bera test for normality [8]. Welch's T-test [9] checks, whether the neighborhood sphere and its hull have the same Gaussian distribution. The resulting mean μ and standard deviation σ are used to modulate the opacity of samples and their shading, so the impact of noise is reduced. The user defines 2D transfer functions in the μ versus σ space. This improves the classification of materials or tissues during volume rendering. The drawback here lies in the fixed distribution model that is fitted to the data. In our approach we do not assume any distribution model, relying instead on empirical distribution functions and statistical tests on them. These functions approximate the underlying data-value distribution.

Praßni *et al.* [10] investigate the usage of shape information in the classification. They represent an object-of-interest in the data with a curve skeleton. The object is split into skeleton regions by corresponding skeleton segments. Tubiness, surfaceness, and blobbiness are used as additional dimensions for the classification of regions. In contrast to computing such spatial properties, we propose to collect and statistically compare data values in regions that are similar to the skeleton regions.

Comparison of regions is also utilized in segmentation(-editing) techniques, which delineate the object-of-interest from the rest of the data. For example, results of the comparison may identify segmentation defects. Our technique [11] detects segmentation defects by analyzing dissimilarities in data values. The object, defined with the segmentation mask, is represented as a set of regions aligned with the skeleton of the object. Inside each such region a histogram collects the data values. The histogram is compared to those from adjacent regions in order to reveal data dissimilarities, associated with segmentation defects. However, this approach cannot

tell whether a certain region is actually part of the object. It just delineates the object from the defects, having detected certain dissimilarities. We provide this missing functionality with our framework and guide the user towards regions that are relevant for editing. This application scenario will be discussed in Section 5.2.

Some methods are designed for temporal data exploration. Among them is the technique of Hochheiser and Shneiderman [12] that enables the user to filter temporal data, composed of numerous time series, with search constraints. The time series are displayed in a two-dimensional plot. Each constraint is defined by a user-drawn rectangle – a timebox. It only matches time series which have data values inside the corresponding rectangle. Having specified several timeboxes, the user gets an uncluttered view of time series of interest. With our approach we may formulate such a filter as a statistical test of similarity to the data in the timebox. However, we leave such extended functionality for future work.

Buono *et al.* [13] proposes a pattern search for temporal data analysis. It is based on the Euclidean distance between corresponding values of two compared time series. To improve the pattern matching, the distance metric includes four additional transformations, which normalize the compared data. The user gets an overview of all regions matched with the specified pattern. Our method performs a statistical test of similarity of data values in sub-regions of a time series and reports the resulting p-values. As we aggregate regions with similar statistical significance properties, the user can explore the data at different levels of detail.

Buono *et al.* [14] focus on the prediction of a time series using patterns found in the temporal data to make extrapolations into the future. Having constructed the empirical distribution functions of the underlying data, one may simulate future data values with our approach. An implementation, however, would require domain-specific knowledge. We leave such extension for future work.

The work of Bögl *et al.* [15] discusses a Visual Analytics approach to the selection of appropriate models for time series data. The user specifies the model by interactively adjusting parameters and selecting characteristic values in a timeline plot. During this specification, the remaining model parameters are estimated. Then, the method provides the user with an in-depth analysis of the selected model's fitness to the data. Various models are compared by their information content. The user selects the most informative one that later can be used for prediction.

Applied to temporal data, our approach solves a different task. It detects the most similar or the most dissimilar occurrences, compared to the exemplar events. Even though the results of our technique are not directly applicable for prediction, they may be useful for finding causes which lead to certain occurrences with respect to the background information.

3 Method

In this section we provide an in-depth description of our approach. We start with *atomic* regions, which are the basis of our method. Inside each atomic region we construct an *empirical distribution function* of data values. We statistically test dissimilarity between data in the atomic regions and values in a set of *markers*, *i.e.*, user-specified exemplars of regions of interest. P-values of the dissimilarity significance are computed. Unions of atomic regions form *composite* regions, which exhibit either significant or non-significant dissimilarities to the markers. The composite regions form a *hierarchy*, varying in size and representing the data with different levels of detail. To support the data exploration task, we provide the user with a *radial tree view* depicting the hierarchy of the composite regions and the derived p-values. Having linked visualizations of the data and the radial tree view, the user follows our *data-exploration protocol*. To support the user in scenarios with dynamically changing data, we display *global statistics on the p-values*. An overview of our method is given in Figure 1.

3.1 Requirements for Statistical Testing

The major requirement of our approach involves the definition of the *atomic* regions. These regions should reflect domain-specific features of the data. The atomicity of the region means that it should contain samples of only one single feature. In other words, each atomic region belongs to a single feature only, *i.e.*, the data-value distribution is assumed to be identical over the region. The atomicity is usually achieved by making the regions small enough. Yet, each region should contain enough samples to properly reconstruct the underlying data-value distribution. In order to provide exact p-values, we restrict the atomic regions to be non-overlapping. To preserve the adjacency of the data at the level of atomic regions, each region should be connected. An atomic region realization that satisfies all the aforementioned requirements will provide meaningful results in terms of statistics. The composite regions then represent the domain-specific features of the data on different levels of detail. The computed p-values differentiate the features of interest from the rest of the data.

Our work concentrates on the statistical framework that operates on top of the realized atomic regions. We provide the realization of atomic regions for the general case

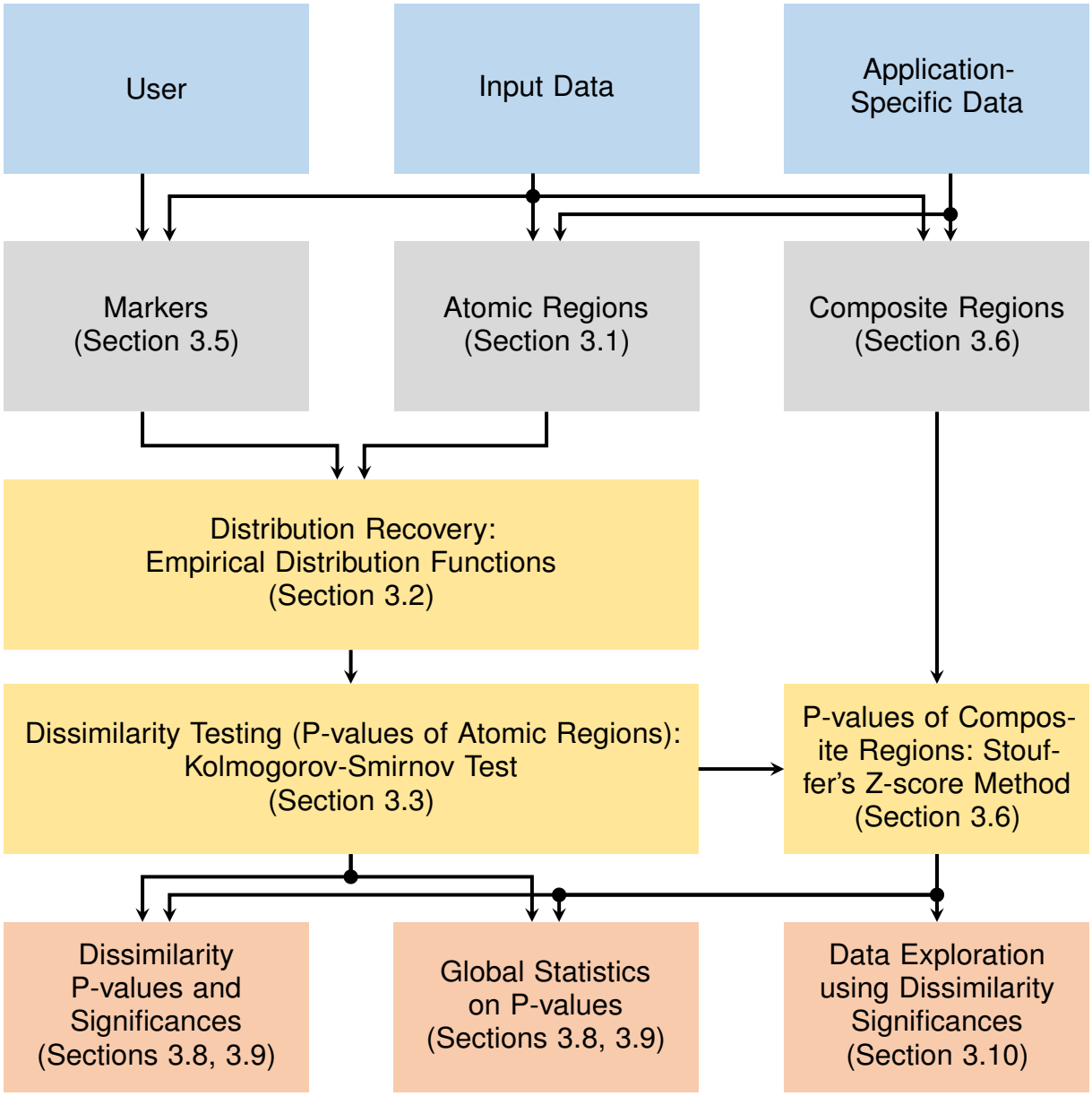

Figure 1: Overview of our statistical method. The information from the data sources is transformed into our internal entities (markers, atomic regions, and composite regions). We recover the underlying data-value distributions in the atomic regions and the markers with empirical distribution functions. Then we perform a statistical test on the significance of dissimilarities between the empirical distribution functions of the regions. We present this information as well as global statistics on p-values to the user. Moreover, we propose a data exploration protocol based on the dissimilarity significance information.

as well as the specialized realization of atomic regions for three-dimensional spatial data. Moreover, we validate these realizations in two application scenarios.

The fundamental property of our framework is its ability to handle stochastic data. Our input data can stem from a spatially and temporally varying probability distribution. For real-world data sources this is generally the case. Natural local variations and inhomogeneities in the data values, acquisition noise, discrete and finite measurement domain and time introduce random fluctuations in the data. We do not assume any knowledge of the underlying distribution and operate on samples from it (*e.g.*, image of a CT (Computed Tomography) scan or temperature measurements in climatic data). Even if the data does not exhibit a stochastic nature, our approach is valid, as we operate on data-value distributions that characterize the various data regions and which are tested for dissimilarity.

In this work we focus on scalar data values. An extension to multi-dimensional data is possible, but we consider the corresponding implementation of the statistical concepts to be future work. The dimensionality of the data domain can be arbitrary though. The density values from medical imaging (CT, MRI) are given on a three-dimensional grid. Abstract time series data consists of samples with one temporal dimension. In the following section we reconstruct the underlying data-value distribution from the samples.

3.2 Distribution Recovery

Given a one-dimensional random variable V and an associated probability function $F(\mathbf{s})$ that varies over an underlying space $\mathbf{s} \in \mathcal{S}$, we take as input data a single outcome $o(\mathbf{d})$ of V on a discrete subset $\mathbf{d} \subseteq \mathcal{D}$ of $\mathcal{S}$. We do not impose any structural constraints on $\mathcal{D}$. In most applications it is isomorph to a structured grid in an m-dimensional space $\mathcal{S}$. Imaging processes, for example, yield two-dimensional and three-dimensional regular grids of data, whereas temporal data is usually taken at equidistant points in time.

The key idea of our approach is to recover the unknown probability function F by collecting outcome data on certain subsets of $\mathcal{D}$ – the atomic regions. We are interested in atomic regions that are similar or dissimilar to the user-specified markers. The similarity is determined by comparing the *Empirical Distribution Functions* (EDFs) that are the distribution functions of the outcome values of V in the regions. We choose the EDFs as we do not assume that the outcome values were drawn from a certain probability distribution (even with unspecified parameters). Instead, we could have selected a distribution model in advance and fitted it to the data. However, this would limit the generality of our method, as the choice of the appropriate distribution model depends on the data modality. Having estimated the parameters of the distribution model from the data, we could not directly apply classical statistical tests [16]. We could have used kernel density estimation (KDE) methods to model the unknown probability density function. This would require a suitable kernel function. Also, the classical statistical tests would need adaptation (*e.g.*, the two-sample shape test on the KDE by Duong *et al.* [17]). With the EDFs we perform the tests directly.

For an atomic region $\mathcal{R} \subset \mathcal{D}$ we build the corresponding EDF $\hat{F}_{\mathcal{R}}$ from the outcomes in $\mathcal{R}$ by

$$\hat{F}_{\mathcal{R}}(t) = \frac{1}{|\mathcal{R}|} \sum_{i=1}^{|\mathcal{R}|} \mathbf{1}\{o(\mathbf{d}_i) \leq t\}, \quad \mathbf{d}_i \in \mathcal{R}, \tag{1}$$

$$\mathbf{1}\{E\} = \begin{cases} 1, & \text{if condition } E \text{ is true} \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

where $|\mathcal{R}|$ is the cardinality of $\mathcal{R}$. Figure 2 illustrates this concept. In order to construct the EDF, Equation 1 requires that the values $o(\mathbf{d}_i)$ are sampled from identical and mutually independent random variables that correspond to the sampling locations $\mathbf{d} \subseteq \mathcal{D}$. The atomic region definition already guarantees that the random variables have the *identical* distribution function $F_{\mathcal{R}}$ (Section 3.1). A set of random variables Ω is *mutually independent* if the probability of any certain outcome of one variable does not depend on outcomes of the remaining variables. Formally, this means that $\forall\{A_1, \ldots, A_M\} \subseteq \Omega$ and $\forall\{a_1, \ldots, a_M\} \in \mathbb{R}^M$ the following statement holds true:

$$Pr\left(\bigcap_{i=1}^{M} \{A_i \leq a_i\}\right) = \prod_{i=1}^{M} Pr(\{A_i \leq a_i\}). \tag{3}$$

This condition is satisfied in the majority of data acquisition settings. For example, in medical imaging data certain physical properties of tissues are measured, such as density or echogenicity. These properties are measured locally at each sampling location independent of the data values from neighboring sampling locations. The independence could be violated due to correlation in the acquisition noise. However, we found this effect to be negligible for our use-cases. As a result, the required EDF is available at each atomic region and can be used for dissimilarity detection.

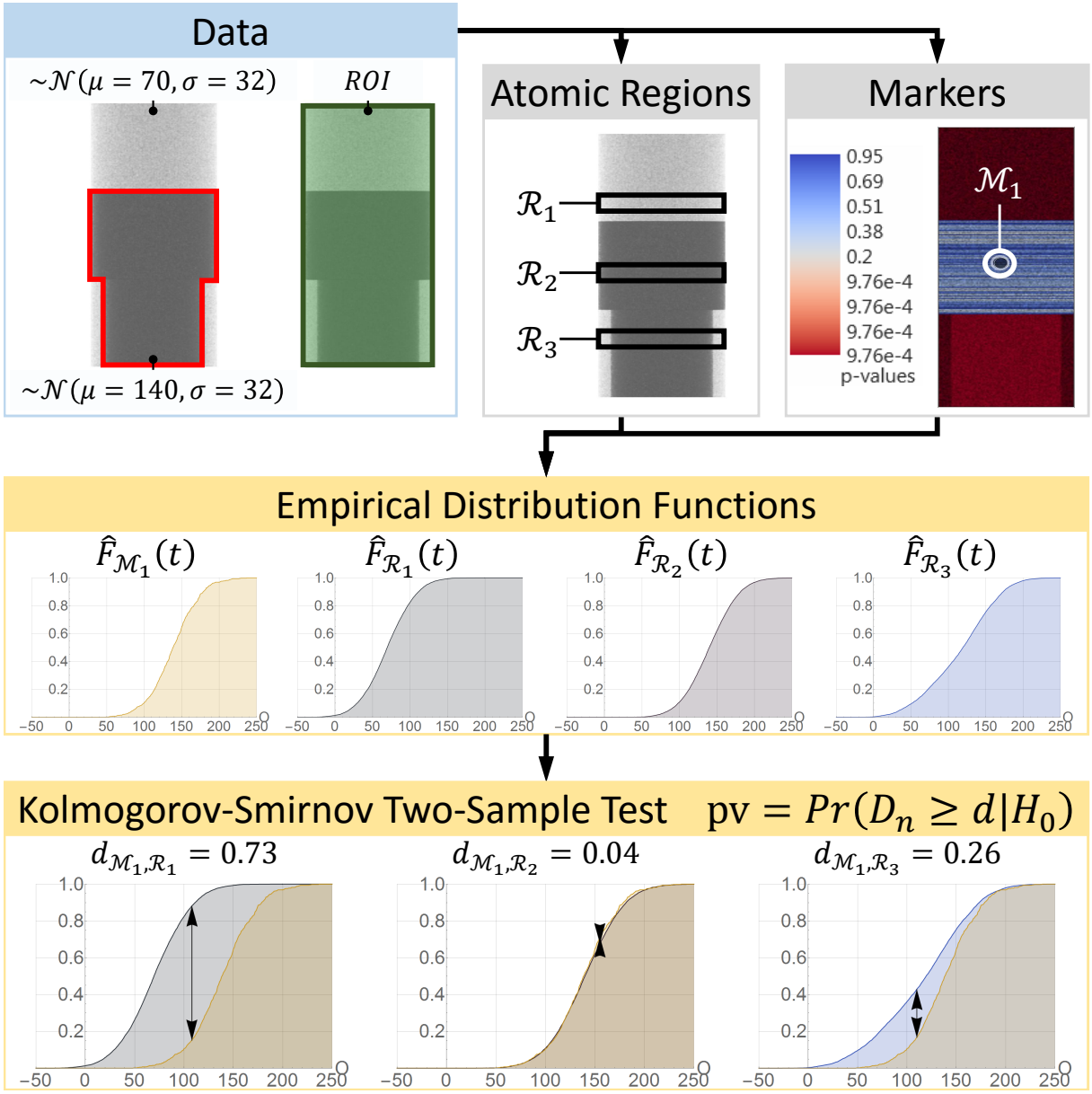

Figure 2: Our method applied to a three-dimensional phantom dataset. The data includes an actual object (red) in the region of interest (green), where the user would like to evaluate the dissimilarities. The region of interest is split into the atomic regions $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3, \ldots$, according to the realization that will be explained in Section 5.2. Once the user has specified a single marker $\mathcal{M}_1$, our method builds empirical distribution functions of the data values in the marker and the atomic regions. Our method employs the Kolmogorov-Smirnov test to compute the p-values of the dissimilarity significance between the atomic regions and the marker.

3.3 Dissimilarity Testing

To quantify the data dissimilarities in two non-overlapping regions $\mathcal{R}_1$ and $\mathcal{R}_2$, we formulate the problem as a statistical hypothesis on the similarity of their respective EDFs. The null hypothesis H_0 is that both EDFs $\hat{F}_{\mathcal{R}_1}$ and $\hat{F}_{\mathcal{R}_2}$ follow the same unspecified distribution. In order to detect a possible violation of H_0 (distribution shape), we employ the *two-sample Kolmogorov-Smirnov test* (KS) with *statistic* D_n [18, 19]:

$$D_n = \sup_{t \in \mathbb{R}} |\hat{F}_{\mathcal{R}_1}(t) - \hat{F}_{\mathcal{R}_2}(t)|, \tag{4}$$

$$n = \frac{|\mathcal{R}_1||\mathcal{R}_2|}{|\mathcal{R}_1| + |\mathcal{R}_2|}, \tag{5}$$

which is a random variable. We test the null hypothesis H_0 by computing the p-value pv as the conditional probability:

$$\text{pv} = Pr\,(D_n \geq d \,|\, H_0)\,, \tag{6}$$

where d is a value of the statistic D_n, computed with the EDFs $\hat{F}_{\mathcal{R}_1}$ and $\hat{F}_{\mathcal{R}_2}$ of the outcome values $o(\mathbf{d}_i)$. This statistical test is illustrated in Figure 2.

In common scientific practice, a *significance level* α is chosen a-priori. The null hypothesis H_0 is rejected if $\text{pv} \leq \alpha$. Here, α determines the probability of falsely rejecting H_0. It is usually set to 0.05, however, it may be different in various scientific domains. In our approach we do not employ a single global significance level, but directly report the p-values to the user. This allows an inspection of the dissimilarity significance of all tested region pairs.

Generally, computing the Kolmogorov distribution D_n is complicated due to a lack of closed-form solutions. Existing approaches often operate with only a one-sample test, where the EDF $\hat{F}_{\mathcal{R}_1}$ is matched against a fully specified distribution. However, we require the more general two-sample test which checks whether two samples were drawn from the same unspecified distribution.

The KS test requires that the hypothetical distribution is continuous, and, therefore, there are no ties in the sampled data (no same values). As we do not assume any restriction on the hypothetical distribution, it can be either continuous or discrete. If the hypothetical distribution is discrete, p-values reported by the KS test are inaccurate [20]. Also, if ties exist in the sampled data, the test yields inaccurate p-values. For generality, we allow the ties.

As stated by Arnold and Emerson [20], in case of a two-sample test with a discrete hypothetical distribution, the distribution of the test statistic depends on this unspecified distribution. In such a situation, a resampling can relax the requirements of continuity and no ties in the data, so the statistical tests can be performed. This was demonstrated by Dufour and Farhat [21] for the KS test.

The resampling of the outcome values $o = \{o_1, \ldots, o_M\}$ is achieved via bootstrap, permutation, or randomized (Monte Carlo) tests. Let us assume the following $\mathcal{E}$ new samples of the outcome values are drawn: $r_1, \ldots, r_\mathcal{E}$. Each new sample contains the same number of the outcome values as the original sample o. The permutation test generates all possible samples without replacement out of o. The bootstrap test iterates over all possible samples with replacement from o. The randomized tests simulate $\mathcal{E}$ samples (with or without replacement) from o. Then the test statistic is computed on the original sample (d_o) and the new samples ($d_{r_1}, \ldots, d_{r_\mathcal{E}}$). Assuming that D_n is a random variable of the test statistic value, the p-value of the test is calculated as follows:

$$\text{pv} = Pr\,(D_n \geq d_o \,|\, H_0) = \frac{1 + \sum_{i=1}^{\mathcal{E}} \mathbf{1}\{d_{r_i} \geq d_o\}}{1 + \mathcal{E}}. \tag{7}$$

The permutation and bootstrap tests have the same asymptomatic power for the KS test, according to Praestgaard [22], and report true p-values. However, taking into account all possible samples requires enormous computational efforts. To alleviate this issue, the randomized tests probe only some randomly drawn samples, and converge to true p-values given enough samples. According to Manly [23], the smallest recommended number of samples is 1000 for the significance level $\alpha = 0.05$, however, Jackson and Somers [24] propose a minimum of $\mathcal{E} = 10000$. We choose the randomized permutation KS test with at least 1000 samples. Next, we discuss alternative tests. Then, we continue with the comparison of data from the atomic regions with the reference values in the markers.

3.4 Alternative Statistical Tests for Dissimilarity

Alternative two-sample distribution shape tests include the Chi-Squared test, Cramér-von Mises (CvM) test, and Anderson-Darling (AD) test. The Chi-Squared test uses binned data. We assume that the data values from the tested regions $\mathcal{R}_1$ and $\mathcal{R}_2$ are binned into the following N_B bins $\mathcal{H}[1], \ldots, \mathcal{H}[N_B]$. The l-th bin counts we denote as $\mathcal{H}_{\mathcal{R}_1}[l]$ and $\mathcal{H}_{\mathcal{R}_2}[l]$ respectively. The binning is usually done in such a way that each bin contains enough samples (more than five). The optimal binning depends on the unspecified hypothetical distribution. The test uses the statistic X [25][pp. 616–617] that follows

a Chi-Squared distribution with N_B degrees of freedom under the null hypothesis:

$$X = \sum_{l=1}^{N_B} \frac{(K_1 * \mathcal{H}_{\mathcal{R}_1}[l] - K_2 * \mathcal{H}_{\mathcal{R}_2}[l])^2}{\mathcal{H}_{\mathcal{R}_1}[l] + \mathcal{H}_{\mathcal{R}_2}[l]}, \quad (8)$$

$$K_1 = \sqrt{\frac{|\mathcal{R}_2|}{|\mathcal{R}_1|}}, \quad (9)$$

$$K_2 = \sqrt{\frac{|\mathcal{R}_1|}{|\mathcal{R}_2|}}. \quad (10)$$

The test reports the p-value for the null hypothesis H_0. However, the Chi-Squared test does not suit our purpose well due to the necessity of data binning.

The KS, the CvM and the AD tests assume a continuous hypothetical distribution. The CvM test uses the statistic W^2 [26]:

$$W^2 = K_3 \int_{-\infty}^{\infty} [\hat{F}_{\mathcal{R}_1}(t) - \hat{F}_{\mathcal{R}_2}(t)]^2 \, \mathrm{d}\hat{F}_{\mathcal{R}_1+\mathcal{R}_2}(t), \quad (11)$$

$$\hat{F}_{\mathcal{R}_1+\mathcal{R}_2}(t) = \frac{|\mathcal{R}_1|\hat{F}_{\mathcal{R}_1}(t) + |\mathcal{R}_2|\hat{F}_{\mathcal{R}_2}(t)}{|\mathcal{R}_1| + |\mathcal{R}_2|}, \quad (12)$$

$$K_3 = \frac{|\mathcal{R}_1||\mathcal{R}_2|}{|\mathcal{R}_1| + |\mathcal{R}_2|}. \quad (13)$$

The AD test computes the statistic A^2 [27]:

$$A^2 = K_3 \int_{-\infty}^{\infty} \omega(t)[\hat{F}_{\mathcal{R}_1}(t) - \hat{F}_{\mathcal{R}_2}(t)]^2 \, \mathrm{d}\hat{F}_{\mathcal{R}_1+\mathcal{R}_2}(t), \quad (14)$$

$$\omega(t) = \frac{1}{\hat{F}_{\mathcal{R}_1+\mathcal{R}_2}(t)[1 - \hat{F}_{\mathcal{R}_1+\mathcal{R}_2}(t)]}. \quad (15)$$

All three statistics D_n, W^2, A^2 can be used in the permutation, bootstrap, and randomized tests in case of a discrete hypothetical distribution or ties in the sampled data. The comparison of the randomized versions of the KS and the CvM tests, conducted by Dufour and Farhat [21], shows little difference in the power of these tests. The authors state that the KS test is more conservative in rejecting the null hypothesis H_0 than the CvM test. In case of a discrete hypothetical distribution, the latter rejects the null hypothesis H_0 more often than the specified significance level. The AD test exhibits a better sensitivity to differences in the tails of the distributions than the KS test, as it is using a weighting factor $\omega(t)$. Feigelson and Babu [28] mention the following commonly overlooked restrictions of the KS test: the independence of the two tested samples and applicability to one dimension only. The same restrictions also apply to the CvM and the AD tests. To keep computational efforts low, we favor the KS test.

3.5 Markers

In most applications, the user is interested in the comparison of the regions with certain reference regions, rather than in the comparison of all possible region pairs. This enables data exploration tasks where reference regions are selected and the most dissimilar (objective O1) or the most similar (objective O2) regions are identified.

As a direct specification of the reference EDF is not possible, we have to estimate it from several reference regions, which we subsequently call *markers*. A marker $\mathcal{M}$ is a connected subset of the sampling space $\mathcal{D}$ of the input data. The EDFs of all the markers approximate the unknown reference distribution function. Given l *disjoint* markers $\mathcal{M}_1, \ldots, \mathcal{M}_l$, we compute the dissimilarity significance for each region $\mathcal{R}_i$ and each marker $\mathcal{M}_j$ by evaluating Equation 4. This gives us a p-value $\mathrm{pv}_{i,j}$ for each such pairing and assigns l p-values to each region. Note that due to their independence requirement, the statistical tests are only valid on disjoint regions. Thus, for each of the aforementioned regions, its intersection with the marker support has to be removed prior to the tests. If the number of samples in the region $\mathcal{R}_i$ is too low for calculating statistics, the corresponding p-values cannot be computed exactly. Denoting the region $\mathcal{R}_i$ without the marker support with $\mathcal{R}_i'$, we check the following criteria:

$$|\mathcal{R}_i'| < S_0, \quad (16)$$

$$\frac{|\mathcal{R}_i'|}{|\mathcal{R}_i|} < S_1, \quad (17)$$

where S_0 is the absolute minimal count and S_1 is the minimal percentage. If any of the criteria is satisfied, we leave the p-values $\mathrm{pv}_{i,1}, \ldots, \mathrm{pv}_{i,l}$ undefined.

Since we are interested in a single p-value pv_i for each region $\mathcal{R}_i$ from all the markers, we combine the p-values $\mathrm{pv}_{i,1}, \ldots, \mathrm{pv}_{i,l}$ of the KS tests. We formulate the task as follows: combine p-values $\mathrm{pv}_1, \ldots, \mathrm{pv}_l$ into a **compound p-value** pv. Nichols *et al.* [29] suggest to combine the individual null hypotheses $H_{0(1)}, \ldots, H_{0(l)}$ with a logical conjunction. The compound null hypothesis $H_0 = H_{0(1)} \wedge \ldots \wedge H_{0(l)}$ is then tested. The corresponding compound p-value pv is computed as follows: $\mathrm{pv} = \max(\mathrm{pv}_1, \ldots, \mathrm{pv}_l)$. This is based on the following logic: the compound p-value is significant only if each individual p-value is significant. The region has significant dissimilarity only if it is significantly dissimilar to all the markers. For each region $\mathcal{R}_i$ we compute its p-value pv_i as

$$\mathrm{pv}_i = \max(\mathrm{pv}_{i,1}, \ldots, \mathrm{pv}_{i,l}). \quad (18)$$

If the region $\mathcal{R}_i$ has insufficient cardinality, then none of the p-values $\mathrm{pv}_{i,1}, \ldots, \mathrm{pv}_{i,l}$ is defined. In this case,

we do not define the compound p-value pv_i as well. In the following exposition, we introduce the concept of composite regions for the data representation.

3.6 Composite Regions

In practice, the number of locations $\mathbf{d}_i$ at which the outcome values $o(\mathbf{d}_i)$ are sampled is usually large. Common medical or biological imaging output, for example, is given on regular grids with millions of cells with data values. This results in a large number of regions where we compute dissimilarity significances. Without additional means, the user would have to investigate all of them during the data exploration, which hinders the analysis task.

To alleviate this issue, we combine the p-values of certain regions into a compound p-value of the regions' union. Figure 3 illustrates such a combination. Given a set of k *disjoint* regions $\mathcal{R}_1,\ldots,\mathcal{R}_k$, we state the following compound null hypothesis: the respective EDFs $\hat{F}_{\mathcal{R}_1},\ldots,\hat{F}_{\mathcal{R}_k}$ follow the same distribution as the EDF of the markers $\mathcal{M}_1,\ldots,\mathcal{M}_l$. For each atomic region $\mathcal{R}_i$ we have computed the p-value pv_i (Equation 18). We now combine $\mathrm{pv}_1,\ldots,\mathrm{pv}_k$ into the p-value for the compound null hypothesis. For combining the p-values in case of the multiple markers, we use the logical conjunction scheme. However, it cannot be applied here, because the compound region is supposed to have significant dissimilarity if any of its subregions is significantly dissimilar to the markers. Therefore, we use Stouffer's Z-score method [30] with the statistic Z_S:

$$Z_S = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}}, \tag{19}$$

$$Z_i = \inf\{t \mid \mathrm{pv}_i \leq Pr(\mathcal{N}(0,1) \leq t)\}, \tag{20}$$

which follows a normal distribution $\mathcal{N}(\mu = 0, \sigma = 1)$. Each combined p-value is assumed to be drawn from a normally distributed random variable $Z_i \sim \mathcal{N}(0,1)$ (Equation 20). The combination method requires these random variables to be independent. Therefore, the p-values should come from independent tests, otherwise the resulting p-value is inaccurate. We ensure this requirement by testing only disjoint regions in the KS test.

As the sum of independent normal distributions is a normal distribution itself, a p-value of Z_S can be computed, which represents the p-value for the compound null hypothesis:

$$\mathrm{pv} = Pr(\mathcal{N}(0,1) \leq Z_S). \tag{21}$$

Later, Liptak, Mosteller and Bush introduced weights $\omega_1,\ldots,\omega_k$ for the combined p-values [31, 32]:

$$Z_W = \frac{\sum_{i=1}^{k} \omega_i Z_i}{\sqrt{\omega_1^2 + \ldots + \omega_k^2}}. \tag{22}$$

The test statistic Z_W also follows a normal distribution $\mathcal{N}(0,1)$, so we obtain a p-value for the compound null hypothesis. The choice of the weights ω_i is an open question, *e.g.*, Whitlock suggests to use the degrees of freedom from preceding statistical tests [33]. In our case, the KS test and the Nichols *et al.* [29] combination formula do not have any degrees of freedom, therefore, we cannot assign the weights ($\omega_1 = \ldots = \omega_k = 1$). One minor modification we introduce relates to atomic regions with undefined p-values. If the p-value pv_i is not defined, then we omit it during the test. If all p-values $\mathrm{pv}_1,\ldots,\mathrm{pv}_k$ are not defined, then we do not define the p-value for the compound region as well.

By partitioning the sample locations $\mathcal{D}$ into a set of disjoint *atomic* regions $\mathcal{R}_i$, the p-values of arbitrary regions in $\mathcal{D}$ can be computed as long as such regions are represented as unions of atomic regions. In this sense, the p-values of the atomic regions form a basis from which the p-values of all possible unions can be computed with Stouffer's method. We refer to such unions as *composite* regions $\mathcal{C}_1,\ldots,\mathcal{C}_h$. The p-value of the composite region $\mathcal{C}_j$, spanning over the atomic regions $\mathcal{R}_{j_1},\ldots,\mathcal{R}_{j_k}$, is computed with Equations 20, 21. To facilitate the data exploration, we suggest that the composite regions vary in cardinality and represent the data with different levels of detail. Together the composite regions form a *hierarchy*. Next, we discuss alternative combination methods. Then, we continue with ranking the composite regions.

3.7 Alternative Combination Methods

There are alternative methods to combine p-values $\mathrm{pv}_1,\ldots,\mathrm{pv}_k$ into a p-value pv for the compound null hypothesis. Winkler *et al.* [34] mention the approach of Nichols *et al.* [29], Stouffer's Z-score method [30] and some other approaches. Edgington [35] suggests the following combination formula: $\mathrm{pv} = \mathrm{pv}_1 + \ldots + \mathrm{pv}_k$. Fisher's combined probability test [36] computes the statistic $X = -2\sum_{i=1}^{k} \log \mathrm{pv}_i$ that follows a Chi-Squared distribution with $2k$ degrees of freedom. Friston *et al.* [37] propose to calculate the p-value as $\mathrm{pv} = (\max(\mathrm{pv}_1,\ldots,\mathrm{pv}_k))^k$. Tippett [38] suggests $\mathrm{pv} = \min(\mathrm{pv}_1,\ldots,\mathrm{pv}_k)$ for the p-values combination. A retrospective discussion by Cousins [39] shows that there is no definite opinion on which test is

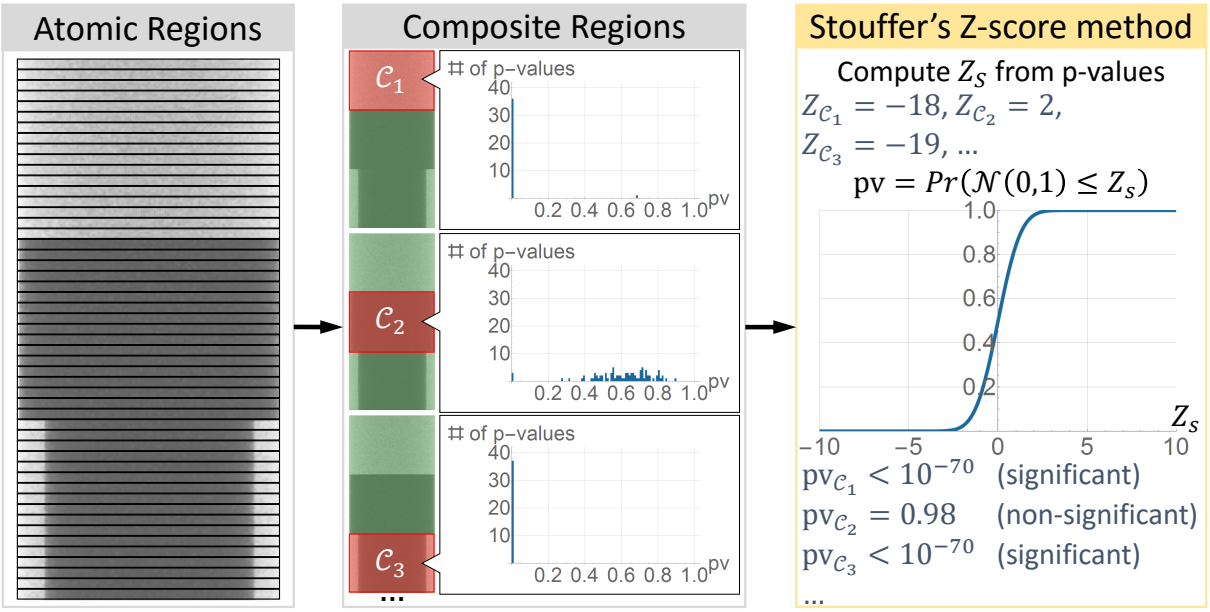

Figure 3: Computation of the p-values for the composite regions of the phantom data from Figure 2. The realization of the composite regions will be discussed in Section 5.2. The composite regions span several atomic regions. We show only three composite regions $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$, but there are more. For these three composite regions we present histograms of the p-values from the underlying atomic regions. For each composite region, we combine these p-values with Stouffer's Z-score method into a single p-value. The resulting p-value refers to the compound null hypothesis H_0 being true for all atomic regions within the composite region.

the best. Different authors recommend Stouffer's Z-score method, Fisher's combined probability test and Tippett's approach. We choose Stouffer's Z-score method.

3.8 Global Ranking by Dissimilarity Significance

In our visualizations we encode the p-values of regions with certain visual attributes. Since the p-values generally span a large range of orders of magnitudes, we cannot directly map them to these attributes. Instead, we rank the regions by their p-values and map the ranks to the visual attributes. Our approach assigns ranks to all the regions, taking into account their cardinalities and preserving their relative differences in p-values. Using the ranks, the user can compare regions with respect to their p-values (Figure 4).

First, we sort all regions by their p-values in ascending order. Next, we put the regions into N_R bins $\mathcal{B}_1, \ldots, \mathcal{B}_{N_R}$. Since the regions vary greatly in cardinality, we equalize the bins by the sum of the cardinalities of the regions, instead of the number of the regions. During the binning process, we keep the order, defined by the p-values. Each bin has its own range of p-values that are determined by its initially assigned regions. However, the p-value ranges can overlap at the end-points. In this case, we resolve ambiguities in bin assignments as follows. If a region has a p-value falling into the ranges of several bins $\mathcal{B}_{k_1}, \ldots, \mathcal{B}_{k_s}$ $(k_1 < \ldots < k_s)$, we put this region into the bin $\mathcal{B}_{k_1}$. The *rank* of a region is the ordinal number i of the bin $\mathcal{B}_i$ it belongs to. Regions with lower ranks have smaller p-values and are more dissimilar to the markers than regions with higher ranks. Depending on the task, the user explores either low ranks (objective O1) or high ranks (objective O2).

Since the composite regions cover several atomic regions, they may have different range of p-values compared to the atomic regions. As we do not compare the atomic and composite regions, they are ranked separately. While the user specifies the markers, we display the ranks of the atomic regions. By looking at the ranks of regions that contain data values similar to the reference ones, the user can quickly validate the employed markers. If these ranks are rather low, then the reference data-value distribution is not adequately approximated by distributions from the markers. The user can refine the employed markers by adjusting their sizes and locations as well as adding new markers. During the data exploration we dis-

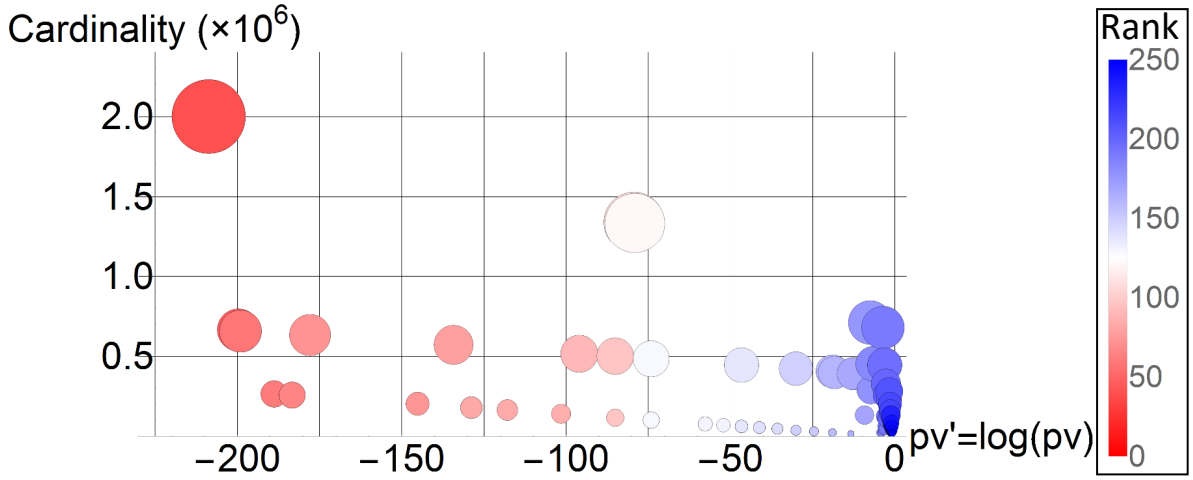

Figure 4: The p-values and the ranks of the composite regions of the phantom data from Figure 2. Each composite region out of 196 is shown with a disk. The area of a disk reflects the cardinality of the corresponding region. The p-values non-uniformly cover the range of $[0, 1]$. The ranks are color-coded. They correspond to the p-values and can be used to compare the regions with respect to the dissimilarity significance. All the ranks are equalized by the sum of the regions' cardinalities (area of the disks) instead of the number of the regions.

play the ranks of the composite regions, aiding the user in efficient localization of the regions with a desired dissimilarity significance. At this point we have established all our abstract concepts. Next, we present the associated visualizations.

3.9 Dissimilarity Significance Visualization

Having collected all necessary statistical information, we now visually encode it to facilitate the data exploration. The composite regions form the hierarchy that represents the data with different levels of detail. We visualize the hierarchy with a *radial tree*, conveying p-values of the regions via color (Figures 5c, 6). We define the level of detail in the hierarchy as the ordinal number of the corresponding level of the radial tree. The levels of detail start from one at the root of the tree and increase. With increasing levels of detail, this visualization occupies successively more screen space. The angular size of a region in the tree is determined by its cardinality. In order to aid the user in reading statistical information, we order the displayed regions with increasing dissimilarity significance in a counterclockwise fashion. Because of this layout, users easily spot relevant regions.

As all the data is represented with different levels of detail, at each such level i there is a region with the most significant dissimilarity $\mathcal{C}^{\max,i}$ and a region with the least significant dissimilarity $\mathcal{C}^{\min,i}$. Such regions are of particular interest to the user, as they fit the selected objective (either O1 or O2) the most. In order to convey the relationships between these regions throughout different levels of detail, we display a path of exploration in the radial tree (Figure 5c2). In case of the objective O1, this path connects regions $\mathcal{C}^{\max,i}$ at each level of detail i. For the objective O2, we connect with the path regions $\mathcal{C}^{\min,i}$. This path connects a composite region to its subregion with the desired dissimilarity significance, *i.e.*, $\mathcal{C}^{\max,i+1} \subset \mathcal{C}^{\max,i}$, $\mathcal{C}^{\min,i+1} \subset \mathcal{C}^{\min,i}$. This way, the user can investigate the data, by adjusting the level of detail interactively. To simplify the exploration of the finer (higher) levels of details, the j-th level parent of the currently selected region becomes a new root of the displayed radial tree. We set the value of j to half of the number of visible levels in the radial tree. Sometimes, the automatically suggested path would lead the user to a region that is sub-optimal in terms of the chosen objective. We denote the level of detail that contains this region as k. In order to proceed, the user employs the radial tree. At the level k the user checks regions, alternative to $\mathcal{C}^{\min,k}$ and $\mathcal{C}^{\max,k}$, and alters the path. In the following sections we explain in detail the data exploration with the radial tree and the paths.

We employ the diverging color map *blue-white-red* by Kenneth [40] for displaying derived statistical information of the regions, as shown in Figures 5c1,d1. The

higher ranks are displayed with blue color tones. This tells the user that these ranks do not carry significant dissimilarities to the markers. White tones indicate an uncertainty related to the middle ranks: their p-values are not small enough to judge if these ranks have a significant dissimilarity. The lower ranks get the user's attention, being highlighted with red tones. The corresponding regions exhibit significant dissimilarities to the markers. The legend is displayed as an annulus around the radial tree, conveying both the diverging color map and the direction of increasing dissimilarity significance. As the data may consist of multiple connected components, at the coarsest level of detail the user can select the preferred component for further exploration. An additional annulus around the tree is divided into sectors that correspond to the connected components, which are given in the data. This annulus is shown only if there are two or more connected components, thus, it is not displayed in Figure 5. We link the data view (Figure 5b) and the radial tree view. This way, they complement each other. The data view displays the data and its regions in two-dimensional screen space. The radial tree view conveys statistical characteristics of the regions and their hierarchy.

The specification of markers requires a certain precision of the user input and involves the selection of samples from the data. Minor variations in the marker's position are allowed as long as it only contains reference data values. If markers belong to two or more features of the data, all such features would be matched. For data with one or two dimensions only the markers can be specified directly in the corresponding data views. In case of three or more dimensions the occlusion problem hinders the exact specification of the region and, thus, we recommend to show such data using the two-dimensional slice view (Figure 5d). We encode the ranks of the atomic regions with the same *blue-white-red* color scheme, as used in the radial tree. This way we convey the dissimilarity significance of the atomic regions with respect to the currently specified markers. While the user specifies the markers, we compute the p-values and ranks interactively to facilitate an immediate check whether the current set of markers adequately approximates the reference data-value distribution.

We support dynamical situations, when the data, the markers, the atomic, or the composite regions change. Such changes can occur even if the analyzed data is spatial. We give as an example an interactive segmentation-editing task, where the segmentation mask is a part of the data. We record in time the consistency that indicates an agreement of data-value distributions between the markers and the rest of the data. As a measure of the consistency, we use the median p-value across all regions. The consistency is communicated to the user via the *p-values timeline plot* (Figure 5a). In this plot we use a logarithmic scale for the vertical axis. Each time a change occurs, we add to the plot a new time step. In order to better depict the trend of the p-values, we connect the median p-values from the individual time steps with lines. We do not use the largest and the smallest p-values, because in real-world data there can always be outlier regions with significant and non-significant dissimilarities to the markers. We do not average the p-values, as they may differ by orders of magnitude. Although the user might be interested in the p-value distribution, the median p-value is sufficient to judge consistency changes in our use-cases.

The p-values timeline plot is useful in three cases. If too few markers are specified, approximation of the reference distribution by their EDFs is inadequate. This situation is indicated by low median p-values. The user should add more markers to improve the low consistency in this case. After each change the user gets a hint whether the consistency actually improves. Such a hint may be useful in several applications, including tuning of filtering as well as reconstruction algorithm parameters, detection of acquisition artifacts, and abnormalities in the data, segmentation editing. The latter application we will investigate in Section 5.2. Having stored the p-value range of each rank for each time step, we can re-rank all current regions according to the p-value ranges of the selected time step. This enables a rank-based comparison that is often required in the data exploration tasks.

3.10 Data Exploration using Dissimilarity Significance

We propose the following protocol for data exploration, aided by the dissimilarity significance (Figure 7). The user starts the exploration in the radial tree view at the coarsest level of detail. Depending on the task, the user chooses the composite region with the most dissimilarity significance (objective O1) or with the least dissimilarity significance (objective O2). Then, the user follows the automatically suggested path in the radial tree, reaching the finer levels of detail. With the linked data view, the user finds the necessary level of detail along the path.

If the current path (Figure 5c2) leads to a result, that is sub-optimal in terms of the chosen objective, a new path should be selected. To do so, the user finds along the current path a coarse region, which is optimal in terms of the chosen objective, and selects it by clicking. Then, the radial tree view is locked on the selected region. In the radial tree view, the user explores a few

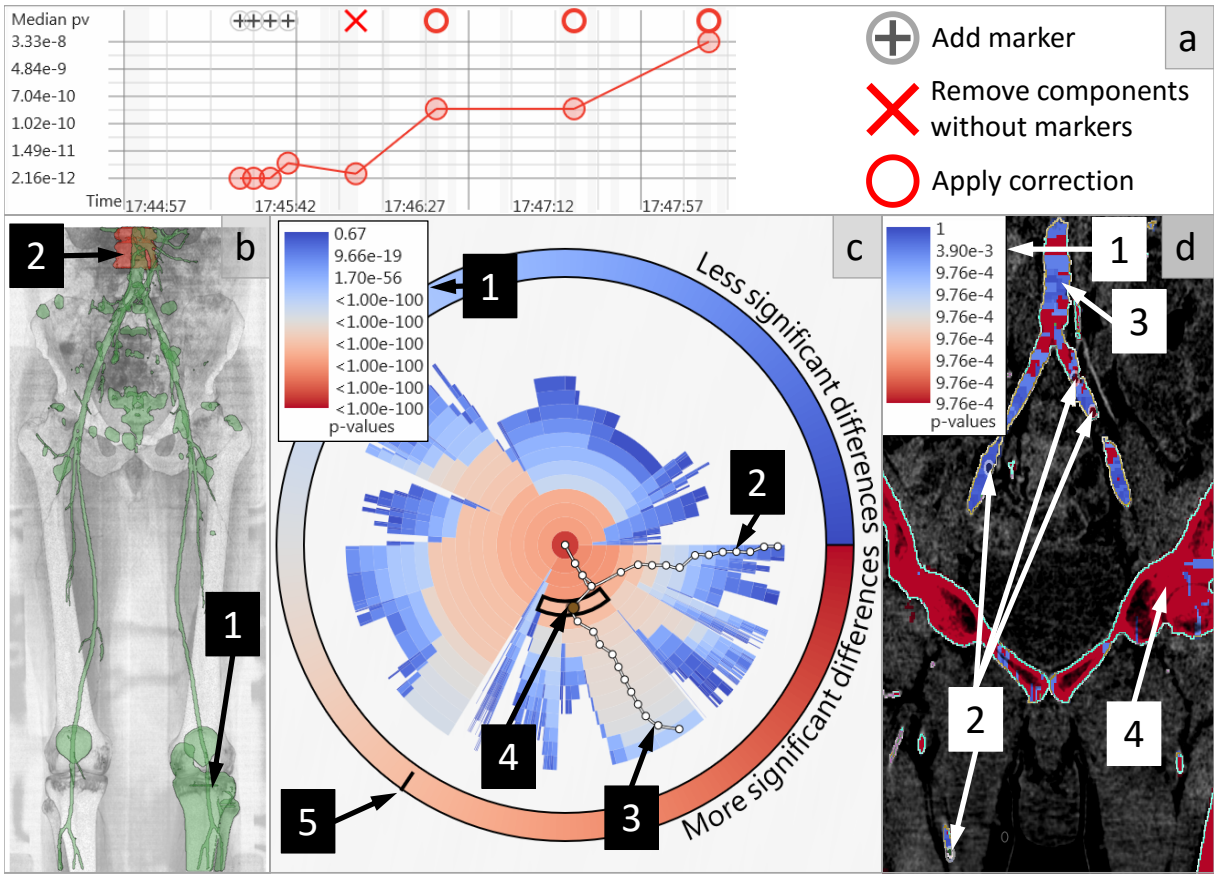

Figure 5: The visualization components of our proposed approach: a) the (median) p-values timeline plot, b) the data view, c) the radial tree view, d) a slice view for marker specification. The timeline plot (a) shows the median p-values after the user specified four markers and applied four operations. The data view (b) shows the entire data (b1, green) and the region of data, which is currently selected in the radial tree view (b2, red). Each composite region is displayed as a node of the radial tree (c). The p-values in the tree are color-coded according to the legend (c1). The user explores the data in the radial tree view by following the paths (c2, c3). The automatically-generated path (c2) links the composite regions with the most significant dissimilarities at each level of detail in the tree. Here, the user chose to explore an alternative path (c3) by selecting composite region (c4). The rank of the selected region is communicated at the annulus (c5). In this application scenario the markers are specified via the slice view (d). The p-values in this view are color-coded according to the legend (d1) The user specified four markers in a vascular structure (d2). The vessels are shown in blue tones (relatively low dissimilarity significance, d3), and the bones are highlighted with red tones (relatively high dissimilarity significance, d4).

subregions of the selected region and selects one that fits the chosen objective the best. The new path, which goes through the selected subregion (Figure 5c4), is automatically generated and displayed (Figure 5c3). Finally, the user continues the data exploration, along the new path.

During the exploration, the user gets immediate visual feedback in both the data view and the radial tree view. This way, the user arrives at the desired region of the data, which has a proper level of detail and fits the chosen objective the best, without inspecting a possibly vast number of composite regions.

4 Implementation

Our implementation uses C# with Intel TBB (C++, on the CPU) and DirectX 11 Compute Shaders (HLSL, on the GPU) for parallelization. We implemented the two-sample randomized permutation KS test on the GPU.

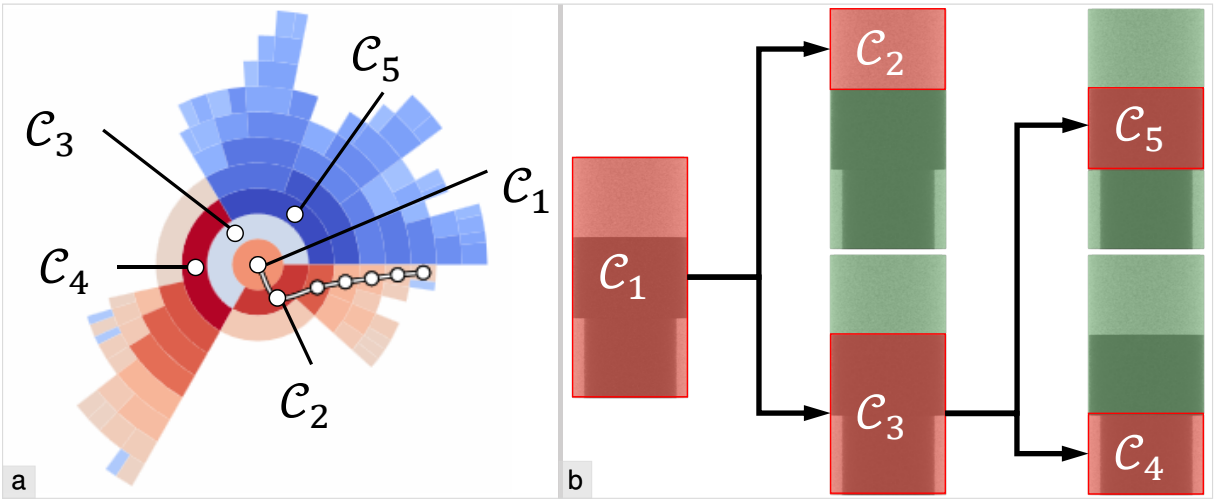

Figure 6: a) The radial tree and some of its nodes. b) The corresponding composite regions (shown in red) of the data from Figure 2. The dissimilarity against the marker $\mathcal{M}_1$ is evaluated. Its significance is color-coded in (a): blue tones depict large p-values (low significance), red tones highlight small p-values (high significance), and white tones are in the middle.

The number of permutations has to be in the interval $[1024, 1024 \cdot 16]$. The KS test reports p-values with single-precision floating-point numbers. We run Stouffer's Z-score method on the CPU in parallel threads. The computation of the normal distribution function is based on works of Cody [41] and Wichura [42]. In order to improve the numerical stability of our statistical framework against overflows and underflows, we compare $\mathrm{pv}' = \log(\mathrm{pv})$ rather than the p-values pv themselves. The logarithm preserves the p-value comparison logic.

We store the samples from the continuous distributions in sorted lists. As for the discrete distributions, the data is organized into histograms, where each bin counts only a single discrete data value.

The value of the parameter $S_0 = 15$ in Equation 16 was determined in the application scenario with small objects (vessels in the human lower extremities), which will be described in Section 5.2. Smaller values of S_0 allow statistical tests on fewer data samples. However, if the data samples are too small, a statistical approach is not appropriate. With $S_1 = 0.5$ in Equation 17, the related criterion is only satisfied in regions that have large overlaps with markers. As data values in the overlaps are excluded from the statistical tests, the remaining parts of such regions are insufficient to perform the statistical tests. However, the users already explored these regions during marker specification, so we can safely exclude them from the further analysis. We use $N_R = 256$ as the number of ranks for implementation convenience, mapping the ranks to the *blue-white-red* color scheme.

5 Results

Our statistical framework supports a general scenario of comparing regions of the data with the user-specified markers. In this section we provide the following components: the basic test for a pair of an atomic region and a marker, the mechanism for combining p-values from multiple markers, and the combination method for composite regions. For each concrete application, one first adapts our framework to the application domain. The adaptation requires a realization of the following abstract concepts: the atomic regions $\mathcal{R}_1, \ldots, \mathcal{R}_k$, the composite regions $\mathcal{C}_1, \ldots, \mathcal{C}_h$, the markers $\mathcal{M}_1, \ldots, \mathcal{M}_l$. Moreover, our null hypothesis H_0 should have a proper interpretation in the application domain.

Atomic regions. We suggest a general atomic-region realization that should be used in case of a lack of a domain-specific realization. All atomic regions have the same cardinality, do not overlap, and each of them represents an m-dimensional hyper-cube in $\mathcal{D}$. The cardinality is application-dependent. It may use natural time intervals of temporal data, *e.g.*, hours, days, months, *etc.* For spatial data it may reflect the expected feature size or units of the volume.

Markers. The user specifies the placements and sizes of the markers. To keep the markers easy to localize, we realize them as m-dimensional hyper-cubes or m-spheres. The markers are usually specified in the analyzed data. However, they can also come from preselected reference data values in external data sources.

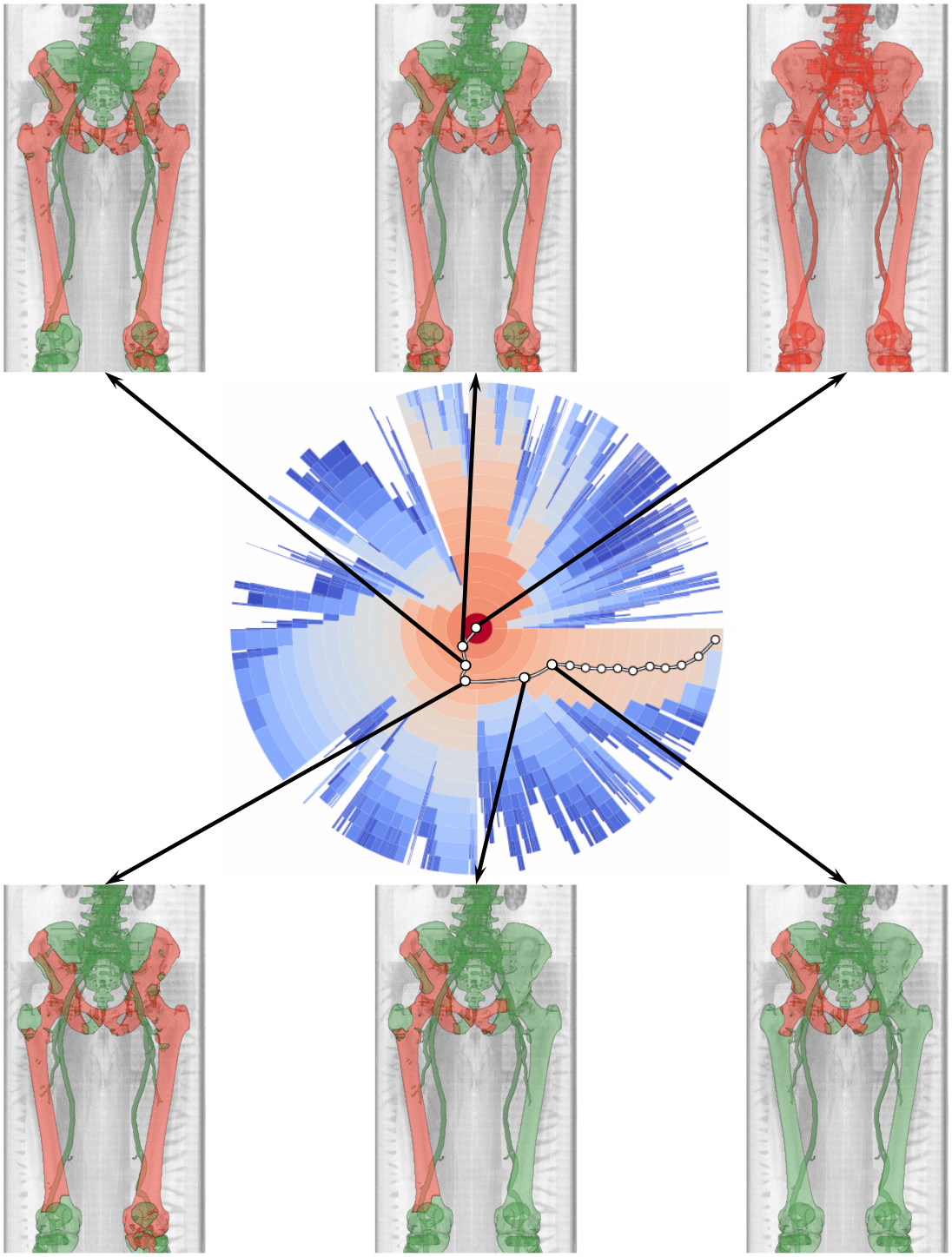

Figure 7: The data exploration protocol, using dissimilarity significance information. The radial tree view shows the data with varying levels of detail. At each level of detail, the region with the most significant dissimilarities is marked with a circle. Such regions across different levels of detail are linked with a path. The first six levels of detail are explored, and the corresponding regions are highlighted in red in the linked data view. The rest of the data is shown in green.

Composite regions. The composite regions can be realized according to the domain-specific logic. If such a realization is missing, we suggest the following general realization. First, we compare the p-values of the atomic regions to the user-specified significance level α. This will effectively remove many atomic regions that are not important for the users. Depending on the objective, we remove the regions with either non-significant (objective O1) or significant (objective O2) dissimilarities. From the remaining atomic regions we construct the composite regions by means of a watershed transformation. For the objective O1 we use the p-value of each atomic region as the height, whereas for the objective O2 we take one minus the p-value. The watershed transformation creates basins and watersheds. The watersheds are the ridges of the height field. They correspond to atomic regions that maximally disagree with the chosen objective compared to their neighborhood excluding the ridges. Each basin contains a local minimum of the height field. Each such minimum corresponds to an atomic region that fits the chosen objective best compared to its neighborhood. The remaining basin elements are separated from other basins by the watersheds. These elements correspond to atomic regions, which fit the chosen objective better than the watersheds. Finally, each basin represents a composite region of the data that fits the chosen objective better than its surrounding. We preserve connectedness of the data between the basins by adding the watersheds to each basin that they separate.

The number of composite regions may be too large for the user to inspect directly. We recommend constructing a hierarchy of the composite regions, if it is not accomplished by the composite region realization. We continue with the hierarchical watershed approach by Hahn *et al.* [43]. For combining the basins, we use the original metric, suggested by the authors. It is equal to the watershed height. Using the metric, we choose two adjacent basins with the lowest watershed between them among all available basin pairs. With the objective O1, we select two basins with the most significant dissimilarities. Under the objective O2, two basins with the least significant dissimilarities are found. We merge these two basins together, creating a new basin (a new composite region). The merging is repeated iteratively until there are no more adjacent basins. The constructed hierarchy represents the data with adjustable levels of detail. Alternatively, one may cluster the composite regions. However, clustering algorithms often require additional input in order to produce sensible results. We favor the hierarchical watershed transformation, as it does not introduce any additional parameters and respects the connectedness of the data.

We demonstrate the generality of our proposed statistical method by applying it to two different types of data. In the following section we describe a temporal-data analysis scenario. Next, our method assists the user during segmentation editing on three-dimensional spatial data from the medical domain.

5.1 Time Series Analysis

In the following we analyze a single time series. We use the general realizations of our abstract concepts in order to investigate it. As a concrete example, we take the maximal daily temperature in Melbourne, Australia in the period 1981–1990 (Figure 8a). The task is to find similar weather conditions (objective O2). The atomic regions span entire months. The user specifies a single marker from 1981-Jun-03 till 1981-Jun-25 (Figure 8b). The hierarchy of composite regions is then constructed and displayed in the radial tree view. The user gets an overview of all found time intervals with possibly similar temperature distributions. In this case, ten intervals are found (Figure 8c): July-August 1981, June-August 1989, June-July 1986, June-July 1982, June-August 1983, June-August 1984, June-August 1990, June-August 1987, June-September 1985, and June-July 1988. They are ordered by decreasing p-values. The user checks each interval (Figure 8f). The first seven time periods have the following p-values: $\mathrm{pv}_1 = 0.388$, $\mathrm{pv}_2 = 0.387$, $\mathrm{pv}_3 = 0.276$, $\mathrm{pv}_4 = 0.198$, $\mathrm{pv}_5 = 0.146$, $\mathrm{pv}_6 = 0.076$, $\mathrm{pv}_7 = 0.057$. These p-values are higher than the set significance level ($\alpha = 0.05$). The user concludes that there is no significant difference so far in the weather conditions between these time intervals and June 1981. The last three time periods have the p-values below the significance level: $\mathrm{pv}_8 = 0.025$, $\mathrm{pv}_9 = 0.023$, $\mathrm{pv}_{10} = 0.014$. The user decides, that the weather conditions exhibit significant differences during these time intervals, compared to June 1981. Time intervals that were discarded during the comparison with the significance level, *e.g.*, September 1983-May 1984, have weather conditions with significant differences compared to June 1981.

5.2 Segmentation Editing

Various applications may benefit from domain-specific realizations of atomic and composite regions. The main idea behind using the specialized realizations is to transfer additional information to the statistical analysis. As an example, we present an adaptation of our method for segmentation editing, where we include spatial features that are captured by skeletonization. Geometric operations on the input segmentation mask define the atomic and composite regions.

Figure 8: A time-series analysis using our dissimilarity-significance method. The task is to find similar (objective O2) weather conditions, considering maximal daily temperature in Melbourne, Australia: a) the measurements throughout the entire observation period (years 1981-1990), b) the user specifies the marker (green) being interested in the temperature distribution from 1981-Jun-03 till 1981-Jun-25, c) the user gets an overview of all found time intervals with possibly similar weather conditions (highlighted with red-white-blue colors), d,e) the p-values legend for the atomic and the composite regions, f) the user investigates ten found time intervals in detail via the radial tree. The first seven time intervals have p-values higher than the set significance level ($\alpha = 0.05$) and indicating no significant difference so far to the weather in June, 1981. The last three time periods have p-values that are lower than the level α and show significant differences in the weather conditions, compared to June, 1981.

Let us assume that we have a three-dimensional data and a mask $\mathcal{L}$ of an object, which is defined as a set of voxels. The mask $\mathcal{L}$, generated (semi-)automatically, often contains regions which, in fact, do not belong to the object. Also, there are actual parts of the object that are not included in the mask $\mathcal{L}$. Both kinds of regions stem from data irregularities/abnormalities and their processing by the segmentation algorithms. The user searches for such regions and either removes or adds them to the mask.

Our statistical approach assists the user during the data exploration process. First, the user specifies the markers that definitely belong to the object. According to our null hypothesis H_0, we identify the first kind of regions as having significant dissimilarities to the markers (objective O1). The second kind of regions, though, exhibits non-significant dissimilarities to the markers (objective O2). Having the dissimilarity-significance information, we suggest the user regions that are relevant for subsequent editing.

The spatial features, captured by the skeletonization, are vital for this scenario, as demonstrated by Karimov et al. [11]. Therefore, we provide the specialized realizations of the atomic and composite regions.

Atomic regions. The concept of the influence zones, introduced by Karimov et al. [44], perfectly matches the atomic regions. The influence zones are constructed as follows. We apply the thinning-based skeletonization by Lee *et al.* [45] on the mask $\mathcal{L}$. During the skeletonization iterations we assign to each peeled voxel the current iteration number. With minor arithmetic operations, these numbers are converted into skeleton distances. We link each voxel in $\mathcal{L}$ to the closest skeleton voxel, building the influence zones. The influence zones cover the mask $\mathcal{L}$ entirely and do not overlap with each other. Each zone can be thought of as a section of the object "orthogonal" to the skeleton. We assume that each influence zone belongs to a single spatial feature of the object if the segmentation mask is correct. Influence zones at segmentation defects, however, may belong to two or more spatial features and exhibit mixed data-value distributions. Such mixed distributions have significant dissimilarities compared to markers, indicating the underlying defects. If the object consists of several tissues or materials, data-value distributions of influence zones should be partially matched with the markers that represent individual tissues or materials. Currently, the partial match is only possible by placing additional markers at interfaces between different tissues or materials. The partial match with reference data-value distributions, generated from groups of weighted markers, is a promising future work. We realize the atomic regions $\mathcal{R}_1, \ldots, \mathcal{R}_k$ as the influence zones, constructed by the algorithms from Karimov et al. [44].

Composite regions. This time we adopt the concepts of the correction regions and the correction operations, introduced by Karimov et al. [11]. The entire mask $\mathcal{L}$ is split into the correction regions according to the histogram dissimilarity metric. The metric analyzes differences between the influence zones and inside individual influence zones in order to find discrepancies, specific to the aforementioned two kinds of regions. To the user this technique suggests a vast variety of correction operations with the corresponding correction regions. The user manually inspects them and selects the ones which actually have to be removed or added to the mask $\mathcal{L}$. The correction regions consist of the influence zones and form a hierarchical representation of the entire mask $\mathcal{L}$. We realize the composite regions $\mathcal{C}_1, \ldots, \mathcal{C}_h$ as the correction regions.

Instead of manually inspecting all correction regions, the user follows the paths in the radial tree view. This way we guide the user to the correction regions that are relevant for the editing task. Each time the user edits the mask $\mathcal{L}$, the p-values timeline plot provides an indication whether the mask $\mathcal{L}$ is improving during the correction process. Non-significant p-values in this plot can be used as a termination criterion of the editing process.

As a concrete example, we perform the segmentation editing of CTA (Computed Tomography Angiography) datasets of the human lower extremities. The goal is to obtain the masks of the vessels, separated from the bones. The given medical datasets exhibit severe pathological conditions as well as major anatomical variability between the patients (missing vessel branches, different branching points). The fact that vascular structures are touching bones causes severe segmentation defects due to the administered contrast agent (Figure 9a). In Figure 9 we present an editing process of one of these datasets. The user starts with the segmentation that contains severe defects, *e.g.*, bones and table fragments (Figure 9a). Four markers are specified inside of the vessels (Figure 9b). In the slice view the user notices, that the vessels are colored with blue tones (relatively low dissimilarity significance), and the bones are mostly shown in red tones (relatively high dissimilarity significance). The user removes the table by removing connected components without markers. The user finds the first three correction operations, following the automatically suggested paths in the radial tree view (Figure 9c-e). This removes major bone structures, but fragments of the vertebrae, the tibia, and the fibula are still present. Along the alternative path the user finds the fourth correction operation that removes the vertebrae fragments (Figure 9f). The remaining bone

fragments are removed by the fifth correction operation, found at the automatically suggested path (Figure 9g). With this the user finishes the editing process, reaching a satisfying result (Figure 9h). For validation purposes, the resulting segmentation is checked against the initial one. The user selects the last step in the timeline plot, where the segmentation is not modified. As expected, all the regions get higher ranks, indicating their relatively low dissimilarity significance (Figure 9j). After editing, the vessel mask has a quality level comparable to the results of a technique used in the daily clinical routine (Figure 9i). All major vessels are preserved, while bones and table fragments have been correctly removed.

5.2.1 Domain Experts Feedback

In order to validate our statistical method with domain experts, we choose the segmentation-editing scenario in the medical image processing domain. Two domain experts evaluated our approach with respect to general usability, benefits of having statistical information available, and interaction aspects. They are radiologists, experienced in segmenting vascular structures from various locations of the human body. They are experienced in correcting segmentation masks, since the results of their currently employed methods do not meet the required quality standards. The resulting vessel mask is used for rendering of vessel reformation images for diagnostic purposes.

During the evaluation session, the domain experts corrected the vessel segmentation masks with our proposed technique T1. Our domain experts also performed the same tasks with the technique T2, described by Karimov et al. [11], as well as with the technique T3 in the AngioVis framework [46] that they use in the clinical routine. The techniques T1 and T2 do not use any domain-specific knowledge, in contrast to the technique T3.

We used ten CTA datasets from the daily routine of our domain experts. First, we showed to the experts the correction of one dataset with our proposed technique T1. The author presented the marker specification and the data exploration protocol to provide a basic understanding. The experts were introduced to the workflow of T1 as well as to the method for a statistical assessment of suggested correction operations. Interaction examples were presented. Then, the experts tried our proposed technique T1 themselves on four other datasets in order to train, as they were completely unfamiliar with it. The same introduction to the technique T2 was made.

Each expert corrected five test datasets with techniques T1, T2, and T3. For the initial segmentation, we employed thresholding with a manually set constant threshold, followed by a few morphological operations and a connected components analysis. The technique T3 used the masks generated by an advanced vessel segmentation technique that distinguishes bones and vessels via an intensity criterion. Thus, the quality of the initial segmentation was superior for the technique T3 than for the techniques T1 and T2. However, the technique T3 and its preceding segmentation method are parts of the well-established clinical workflow, so we did not compensate for this difference in our results. Each domain expert first corrected the test datasets with T1, then processed them with T2, and after three days processed them with T3. After the experts finished their tasks, they answered our questionnaire. We measured the time that was required to correct the vessel masks. On average, with our technique T1 the experts spent five minutes fifty seconds on each dataset. The users never finished their tasks with the technique T2, as even one single dataset was not completely corrected after twenty minutes. With the base-line technique T3 the experts achieved the desired vessel segmentation in six minutes twenty seconds per dataset on average.

The domain experts assessed the usability of our method with the System Usability Score by Brooke [47]. The average score was 86 points out of 100, which approximately corresponds to the 90th percentile of the perceived **usability**.

The experts felt **confident** while performing their tasks, having the statistical information from our proposed method. They consider this information as **useful**. According to our domain experts, the radial tree view significantly improved the **overview** of suggested correction operations compared to the technique T2. Also, the experts liked the **exploration** of the suggested correction operations with the paths, which indicate the statistically most significant dissimilarities. The markers' **specification** via the slice view was rather easy and straightforward to our experts. With the markers the experts were able to **precisely** indicate the regions of interest, *i.e.*, the vascular structures. Our chosen **color-coding** scheme proved to be good for both the overview and exploration purposes, judging by responses of the experts and the achieved segmentation results. Finally, with the help of the timeline plot as an **overview**, the experts were able to see a progressive improvement of segmentation quality during the correction. Details on the scores are available in Figure 10a.

Even after a very short adaption period of ninety minutes, the experienced radiologists liked the radial tree view of our technique T1. It enabled the experts to efficiently navigate between different correction operations. The paths, displayed in the radial tree view, were beneficial for the experts, who used this feature at each editing step. These

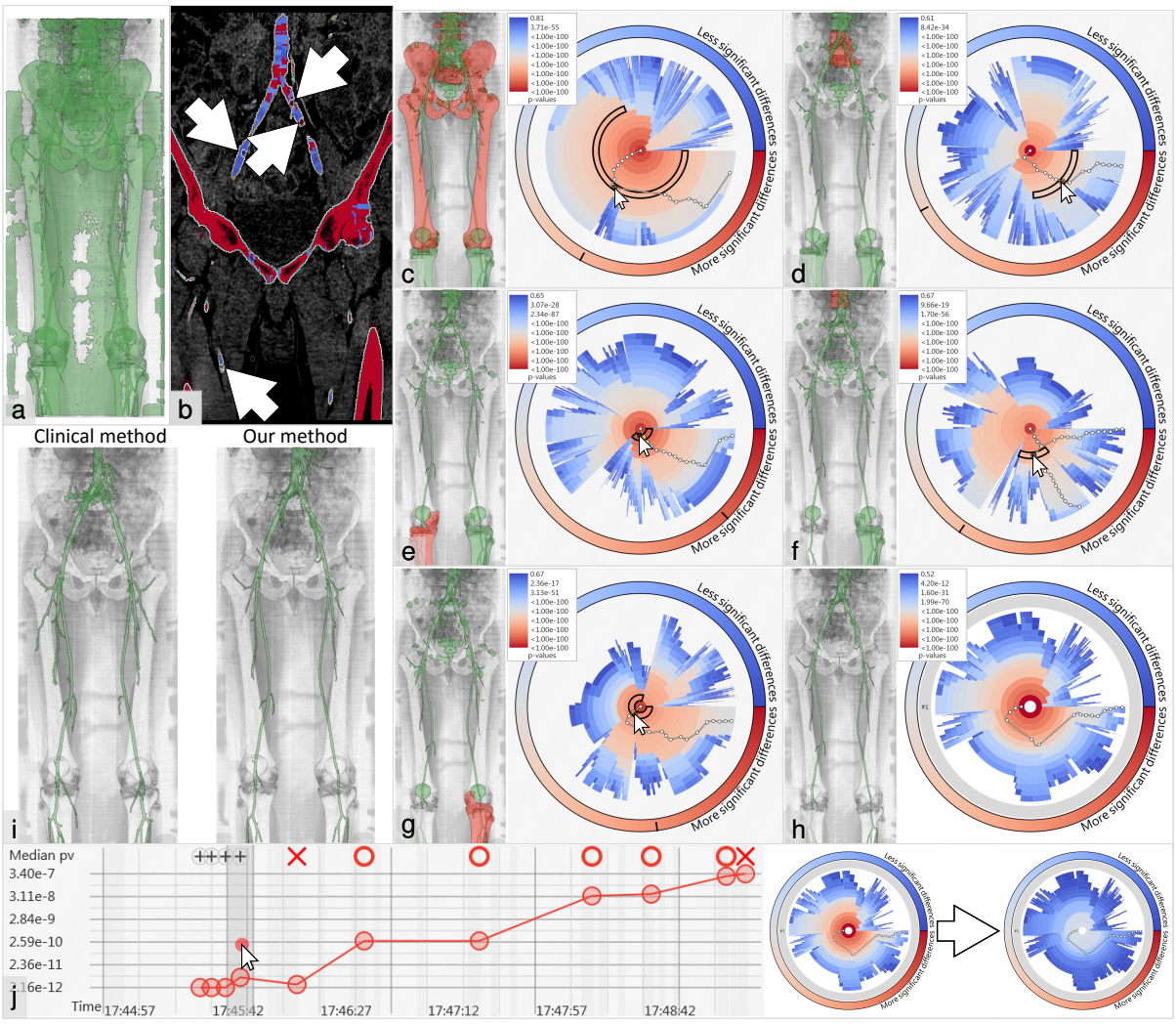

Figure 9: The segmentation editing scenario using our dissimilarity significance method. The task is to edit a segmentation mask of vascular structures in CTA data of human lower extremities. The composite regions represent the mask with different levels of detail. In order to correct the mask, the user selects regions for removal. The user is interested in the regions exhibiting significant differences (objective O1) to the object-of-interest (vessels): a) the automatic vessel segmentation (green) exhibiting defects (fragments of the scanning table and bones are included), b) the user specifies four markers inside of vessels (indicated with arrows), c-e) following the automatically suggested paths, the user finds the first three correction operations (removal of red parts), f) the user follows the alternative path and finds the fourth correction operation (removal of red part), g) the user finds the fifth correction operation, following the automatically suggested path (removal of red part), h) the user achieves a satisfying result after only five operations, i) comparison of our method's result with a result obtained by employing a technique from clinical routine: the results are of the comparable quality, few small vessels are not properly segmented as they are out of interest for the radiologists, j) the user validates the segmentation-quality improvement by comparing the result (h) to the original state (a) (selected in the timeline plot) – the composite regions are finally more in the blue color range and achieve higher ranks, indicating higher p-values than those of the original segmentation mask.

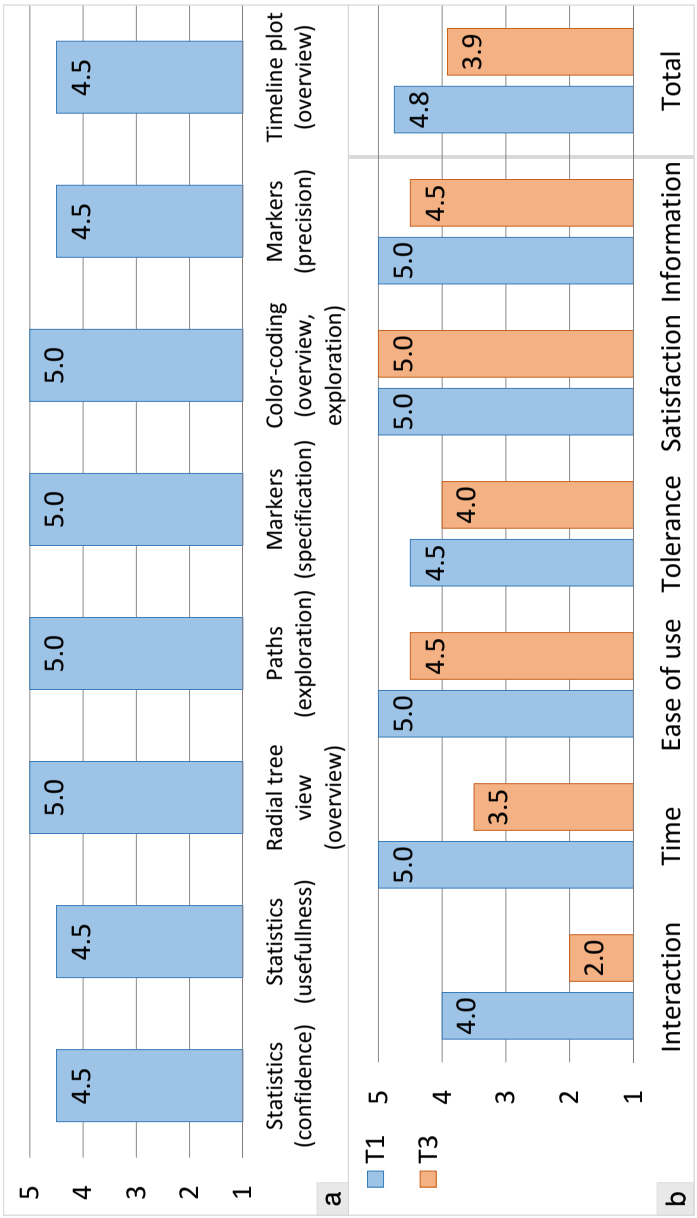

Figure 10: a) Evaluation of our technique. b) Comparison of our method (T1) to the clinical method (T3). Grades range from 1 (worst) to 5 (best).

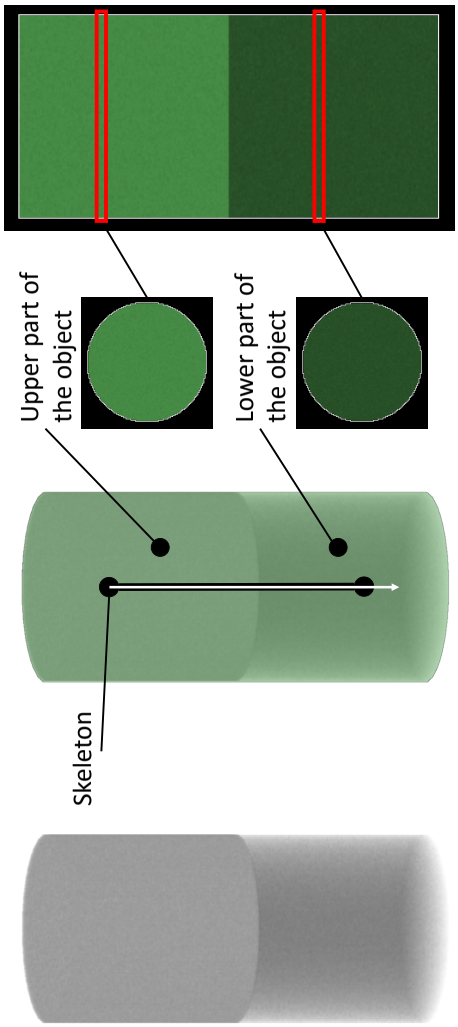

Figure 11: The phantom object, its two parts and its skeleton.

paths, being generated from the statistical information, indicate the usefulness of our proposed statistical dissimilarity evaluation method. The results of the technique T2 strengthen this fact further. Although the correction operation suggestions were completely the same for the techniques T1 and T2, the experts were unable to finish the editing with the technique T2 in a reasonable amount of time (twenty minutes per dataset) due to the data complexity. The main reason was a lack of overview of the correction-operation suggestions, as pointed out by the mediocre score in the evaluation by Karimov et al. [11]. In our technique T1, the radial tree view and the paths, carefully investigated by the radiologists, rectify this deficiency and reduce necessary interaction. We may conclude that our technique T1 enables the exploration of data with complex structures, such as vessels.

Compared to the base-line technique T3, our technique T1 required less **interaction**. The task completion **time** was slightly shorter as well. Both techniques were considered **easy** to use by the experts. The **tolerance** against imprecise input was rated slightly higher for our technique T1. Both techniques T1 and T3 allowed the experts to reach **completely satisfying** results. As for support of the users during the interactive editing with relevant **information**, our technique T1 was ranked slightly better than the base-line technique T3. Detailed information on a comparative evaluation can be found in Figure 10b. In **total**, our technique T1 got a better score than the base-line technique T3. This was achieved by extending the existing technique T2 with our dissimilarity significance computation and novel visualization components. Without these extensions, according to one of our domain experts, the technique T2 "was unable to process complex datasets such as peripheral arteries". Still, the technique T2 properly handles less complex objects, such as single organs or specimens.

5.2.2 Robustness Tests

In order to verify the robustness of our statistical method, we conducted several tests. For these tests we used three-dimensional spatial data and the specialized realizations based on the influence zones and correction regions. We chose it because such kind of data is usually harder to analyze than a time series due to the higher dimensionality.

Robustness of the Statistical Tests

In order to evaluate the **robustness** of the employed **statistical tests** (the KS test and Stouffer's Z-score method), we conducted checks with a phantom 3D object composed of two parts, each exhibiting different data value distributions (six cases). The object is depicted in Figure 11. A single marker was placed in one of two positions: the middle point of the upper part and the middle point of the lower part of the object. The marker radius was 16 voxels. The differences between the distributions of two object parts were varied with delta $\Delta d \in \{0,1,2,4,8,16,32,64\}$. As a sanity check, we included the situation where there is no difference between the distributions (*i.e.*, $\Delta d = 0$). The original parameters were $\mu_1 = 30, \sigma_1 = 10$. The cases were:

- case *C1* - the lower part has the Gaussian distribution $\mathcal{N}(\mu_1, \sigma_1)$, and the upper part has the Gaussian distribution $\mathcal{N}(\mu_1 + \Delta d, \sigma_1)$
- case *C2* - the lower part has the Gaussian distribution $\mathcal{N}(\mu_1, \sigma_1)$, and the upper part has the Gaussian distribution $\mathcal{N}(\mu_1, \sigma_1 + \Delta d)$
- case *C3* - the lower part has the Gaussian distribution $\mathcal{N}(\mu_1, \sigma_1)$, and the upper part has the Poisson distribution $\mathcal{P}(\mu_1 + \Delta d)$
- case *C4* - the lower part has the Gaussian distribution $\mathcal{N}(\mu_1, \sigma_1)$, and the upper part has the distribution $A_1 \sim \mathcal{P}((\sigma_1 + \Delta d)^2) + \mu_1 - (\sigma_1 + \Delta d)^2$ with $E[A_1] = \mu_1$ and $Var(A_1) = (\sigma_1 + \Delta d)^2$
- case *C5* - the lower part has the Poisson distribution $\mathcal{P}(\mu_1)$, and the upper part has the Poisson distribution $\mathcal{P}(\mu_1 + \Delta d)$
- case *C6* - the lower part has the Poisson distribution $\mathcal{P}(\mu_1)$, and the upper part has the distribution $A_2 \sim \mathcal{P}(\mu_1 + 3\Delta d) - 3\Delta d$ with $E[A_2] = \mu_1$ and $Var(A_2) = \mu_1 + 3\Delta d$

Following the common representation of three-dimensional spatial data, the generated sample values were rounded to integral numbers. The data values generated were numbers from -32768 to 32767 (signed 16-bit integers).

The detailed information on the p-values along the skeleton is shown in Figures 12-17. The significance level was $\alpha = 0.01$. We note that the differences are statistically significant starting from $\Delta d = 1$ in the case C1, $\Delta d = 2$ in the case C2, $\Delta d = 0$ in the case C3, $\Delta d = 1$ in the case C4, $\Delta d = 1$ in the case C5, and $\Delta d = 2$ in the case C6. We conclude that our statistical pipeline detects significant dissimilarities in all tested cases, even if the differences are rather small ($\Delta d \geq 2$). If $\Delta d \geq 1$, the p-values of the object part without a marker indicate significant dissimilarities ($pv \leq 0.01$), as illustrated in Figure 18. The p-values of the object part with a marker, however, convey no significant dissimilarities. In case of no differences ($\Delta d = 0$), the reported p-values show

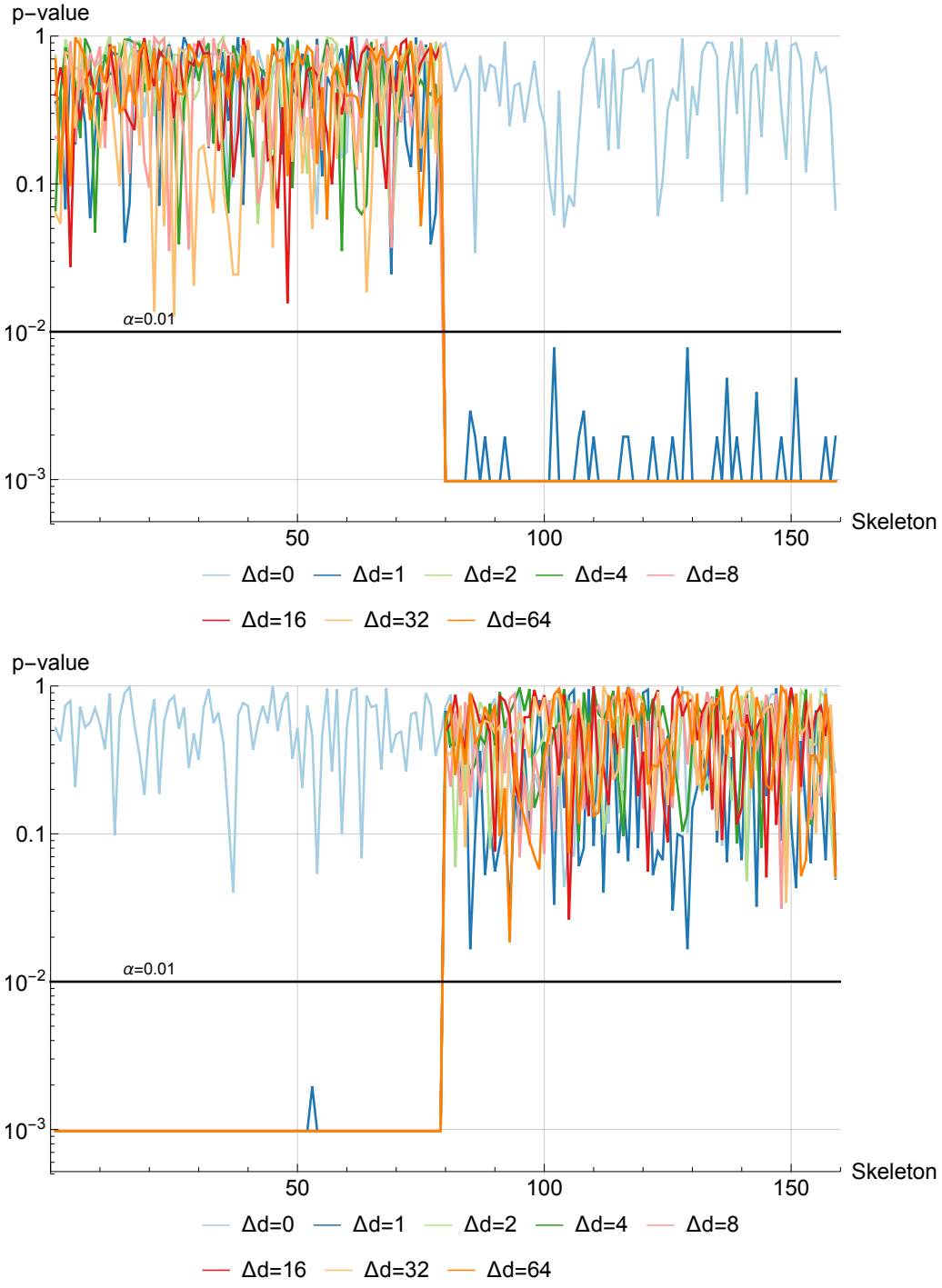

Figure 12: Case C1: p-values along the skeleton with respect to different Δd. The top image shows the p-values if the marker is placed in the upper object part. The bottom image reports the p-values if the marker is located in the lower object part.

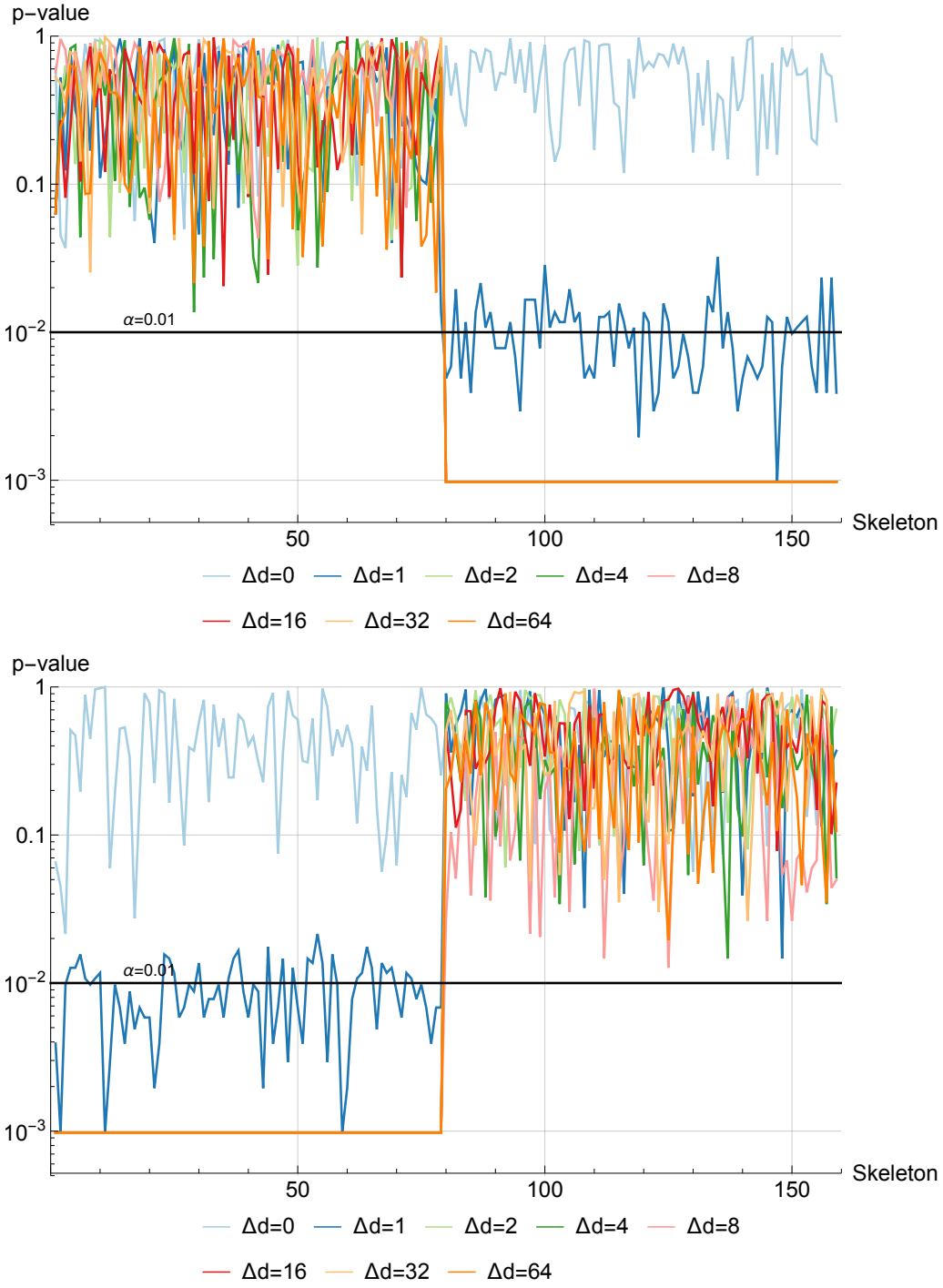

Figure 13: Case C2: p-values along the skeleton with respect to different Δd. The top image shows the p-values if the marker is placed in the upper object part. The bottom image reports the p-values if the marker is located in the lower object part.

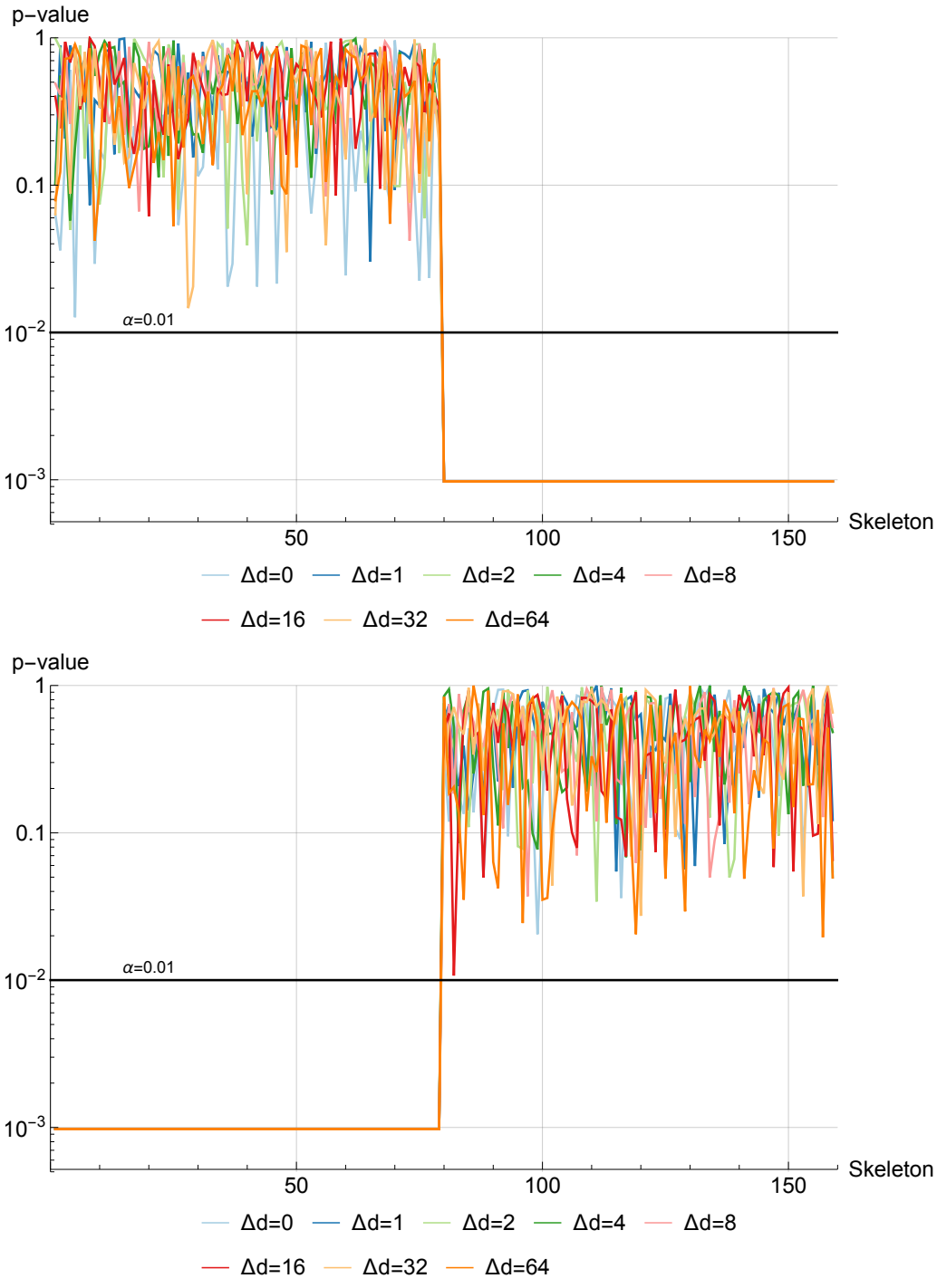

Figure 14: Case C3: p-values along the skeleton with respect to different Δd. The top image shows the p-values if the marker is placed in the upper object part. The bottom image reports the p-values if the marker is located in the lower object part. The significant differences in the case of $\Delta d = 0$ are caused by two different distributions (Gaussian, Poisson), used for the upper and the lower parts of the object respectively.

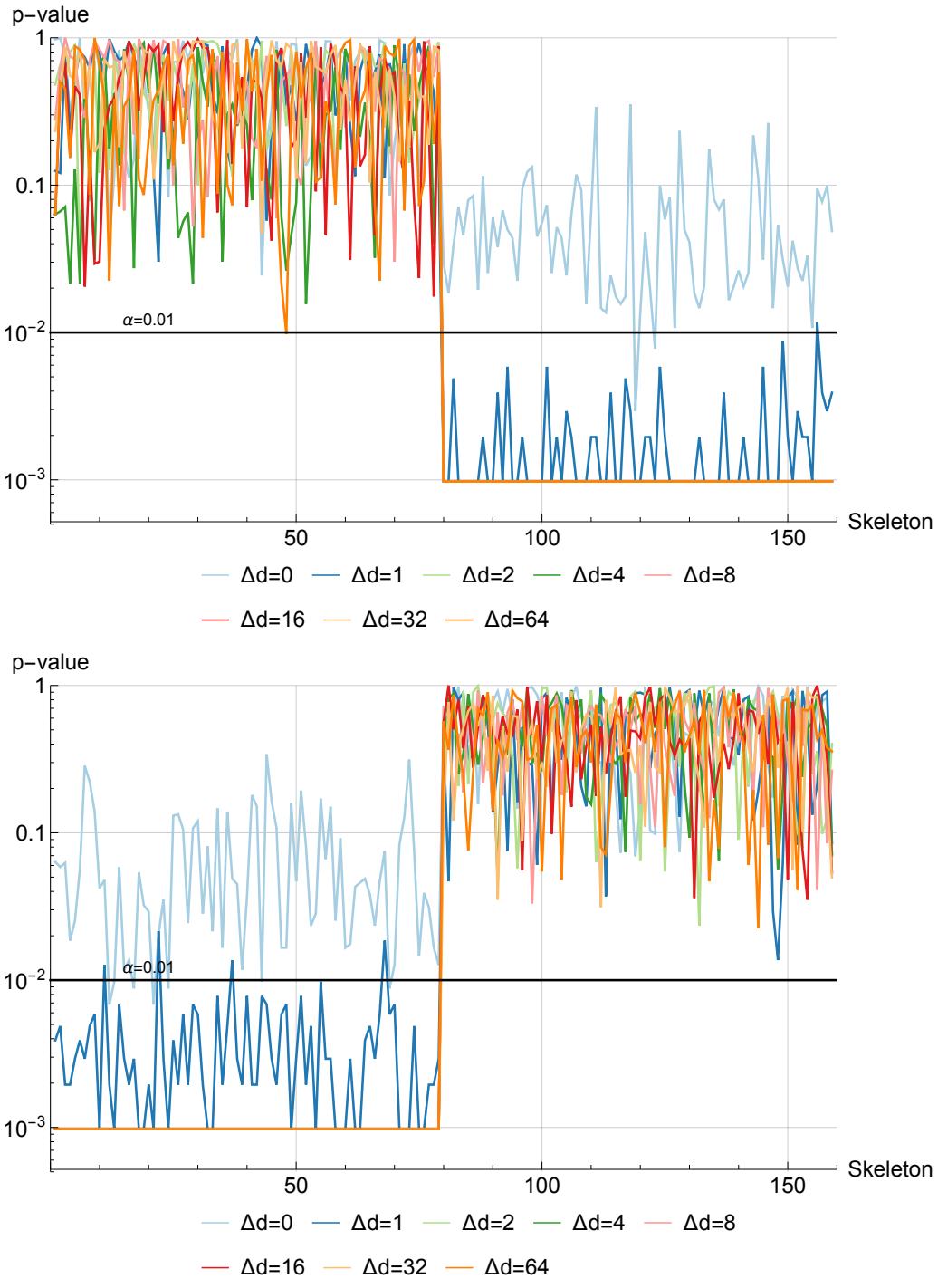

Figure 15: Case C4: p-values along the skeleton with respect to different Δd. The top image shows the p-values if the marker is placed in the upper object part. The bottom image reports the p-values if the marker is located in the lower object part. The significant differences in the case of $\Delta d = 0$ are caused by two different distributions (Gaussian, Poisson), used for the upper and the lower parts of the object respectively.

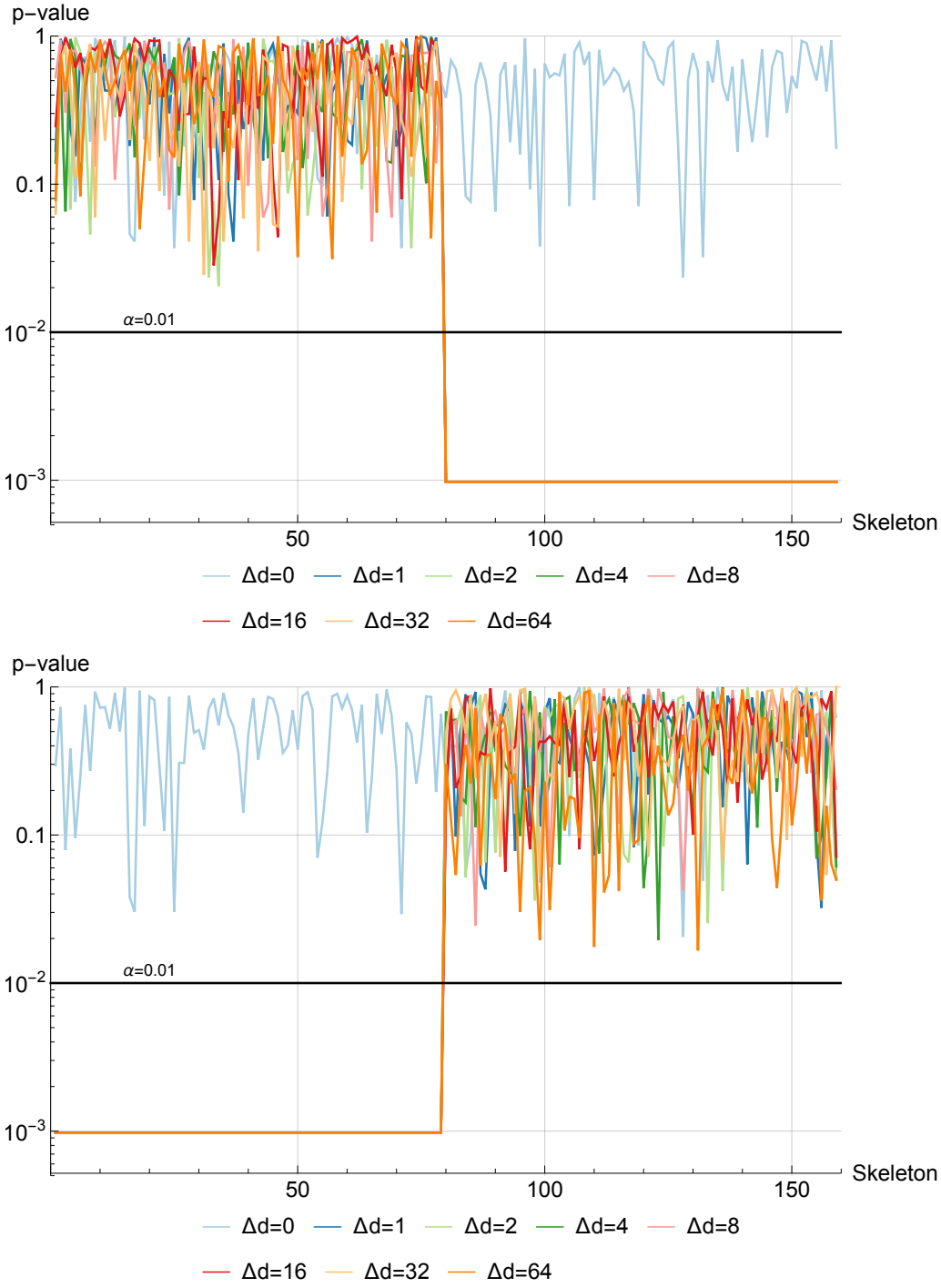

Figure 16: Case C5: p-values along the skeleton with respect to different Δd. The top image shows the p-values if the marker is placed in the upper object part. The bottom image reports the p-values if the marker is located in the lower object part.

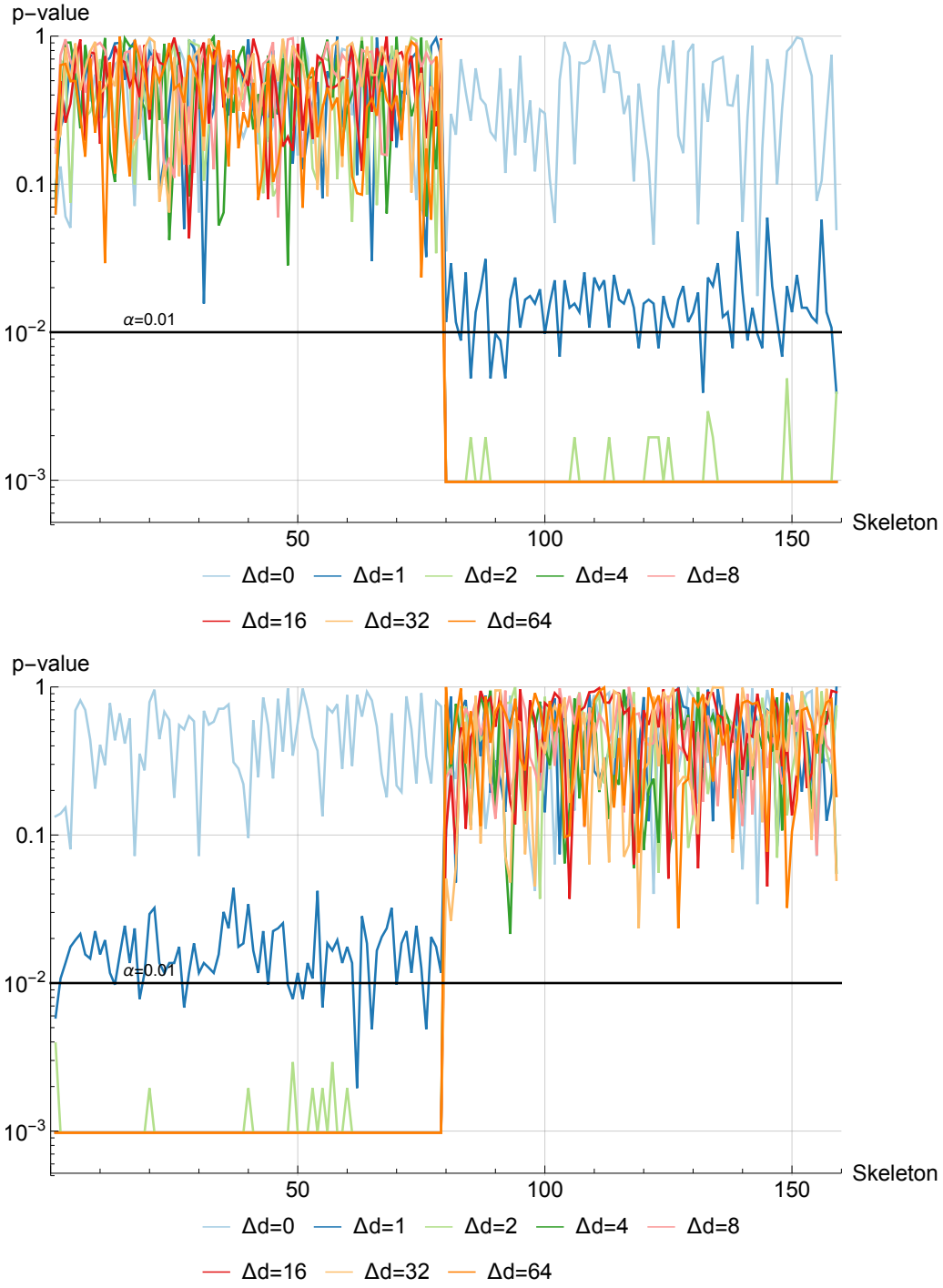

Figure 17: Case C6: p-values along the skeleton with respect to different Δd. The top image shows the p-values if the marker is placed in the upper object part. The bottom image reports the p-values if the marker is located in the lower object part.

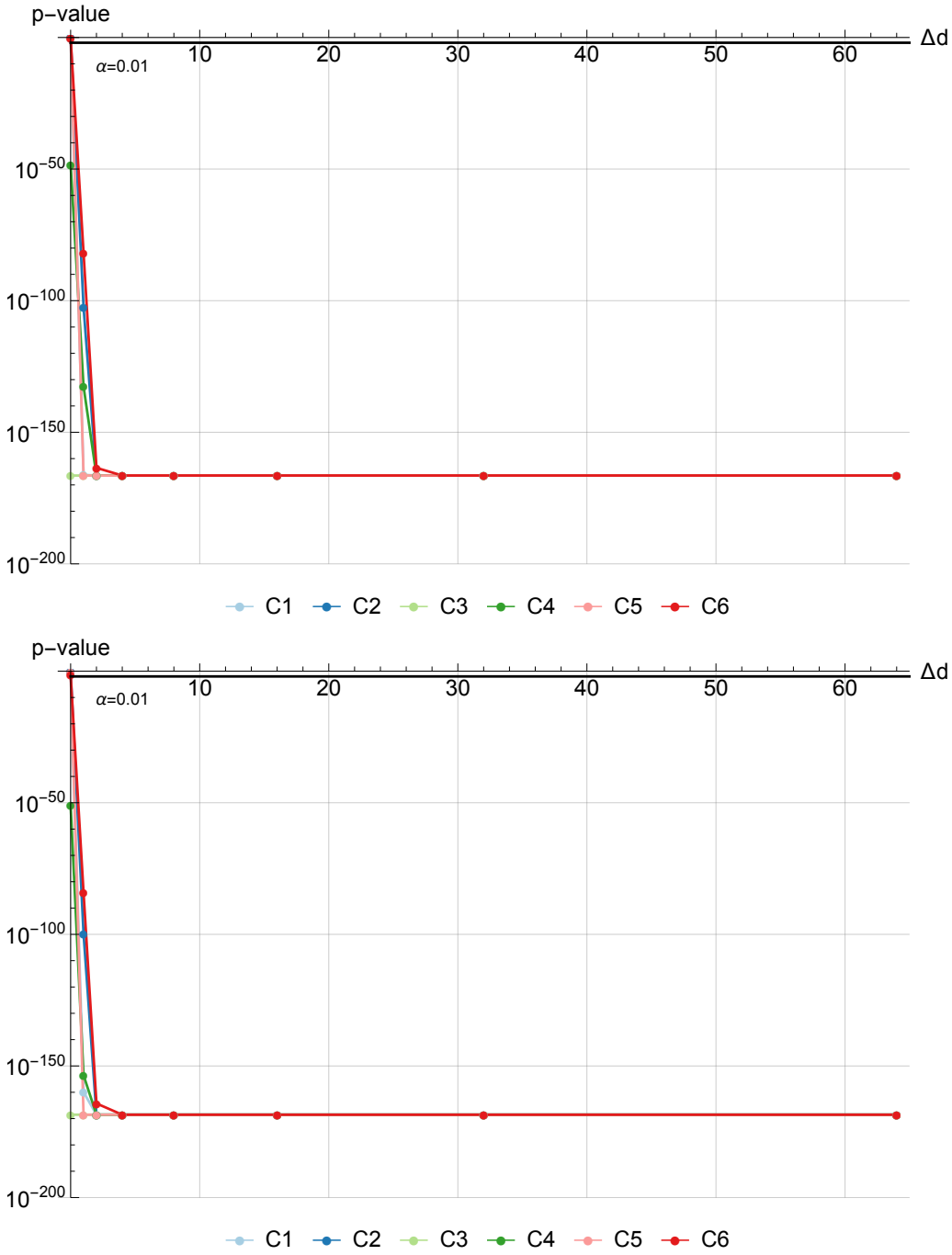

Figure 18: P-values for the upper (top image) and the lower (bottom image) parts of the object with respect to different cases and Δd. The single marker is placed either in the upper part (bottom image) or in the lower part (top image) of the object. At $\Delta d = 0$, the reported p-values are higher than the significance level α in the cases C1, C2, C5, C6, indicating no significant differences. In the cases C3 and C4, however, the p-values are below α. This corresponds to significant differences between two different distributions. With $\Delta d \geq 1$, the reported p-values are also below α. This implies significant differences between distributions.

no significant dissimilarity as well (only applicable for cases C1, C2, C5, C6).

Robustness against Noise

To verify the **robustness** to **noise**, two kinds of artificial noise were added to the medical dataset depicted in Figure 7. The first kind of noise was normally distributed with a mean $\mu = 0$ and varying standard deviation σ according to Table 1. The second kind of noise had a Poisson distribution with a varying mean μ as shown in Table 1. The PSNR (Peak Signal-to-Noise Ratio) was used to evaluate the degree of artificial noise. We used 2048 as the maximal value (upper range for integer values used in CTA data). The segmentation mask was generated anew for each μ and σ. Our statistical method was aiding an interactive segmentation-editing framework. With a thresholding algorithm, masks of bones and vessels were generated. The author then edited them to keep only the vessels. For each dataset, only four markers were specified. In each corrected segmentation mask, the markers were placed at exactly the same positions. The markers' positions were determined in the original dataset without noise. In the test we used one medical CTA dataset of the human lower extremities. The dataset had 800 slices; each slice had resolution of 512×512 pixels.

At each tried noise level, the user achieved the desired result. As a quality measure, we employed the Jaccard coefficient J between the output mask and the reference solution, obtained from the data without the artificial noise and manually refined by one of our domain experts. The Jaccard coefficient is a normalized similarity measure, which reaches one in case of a complete match between two tested sets. The reference solution required three editing steps as well as some voxel level editing, accomplished in 90 seconds. The input masks had a Jaccard coefficient lower than 0.05, indicating a large deviation from the reference solution. With all tried noise models, the editing was successful, as shown by a significant improvement of the quality measure. Moreover, our statistical method was discriminating bones from vessels, as demonstrated by the resulting segmentation masks of the vessels. Detailed information on editing time, number of steps and the output quality can be found in Table 1. For comparison purposes, the resulting masks are shown in Figure 19. The quality of the result, measured with the Jaccard coefficient between the output mask and the reference solution, was above 0.88.

Robustness of the Marker-based Reconstruction of the EDFs

We evaluated how different numbers of markers can reconstruct the complex distribution function of a real-world dataset with the following test. For this purpose, we used the median p-value across all atomic regions as a measure of consistency between the empirical distribution functions of the markers and the underlying data value distribution.

We randomly placed 50 markers inside the vessels of the human lower extremities in a rather challenging medical dataset (Figure 7). To estimate the stability of marker placement, we repeated the procedure 50 times. We set the significance level α to 0.05. In the test we used one CTA medical dataset. The dataset had 800 slices; each slice had resolution of 512×512 pixels. We used the vessel segmentation mask, edited and manually refined by the domain expert.

Resulting median p-values for 50 sets of randomly placed markers are shown in Figure 20. As expected, the more markers we allocate, the more stable is the reconstruction of the distribution function. This is reflected in relatively large median p-values, which became non-significant after five markers in all 50 trials. The more markers we use, the less is the median p-values variation. We may conclude that EDFs from five or more markers reconstruct the data-value distribution of the real-world object with sufficient precision.

5.3 Performance Tests

In order to test the performance of our statistical approach, we used an Intel Core i7-2600K 3.4 GHz CPU with 16 GB of RAM and an NVidia GeForce Titan X GPU. In the case of the temporal data with 3650 samples, the calculation time is 0.004 seconds with one marker, 0.006 seconds with two markers, 0.01 seconds with four markers, and 0.017 seconds with eight markers. For the spatial example we use ten CTA datasets of the human lower extremities with a slice resolution of 512×512 pixels and the number of slices ranging from 700 to 975. The datasets capture vessels and bones. The average calculation time is 0.51 seconds with one marker, 0.72 seconds with two markers, 1.16 seconds with four markers, and 2.05 seconds with eight markers.

6 Discussion and Limitations

We identified certain limitations of our statistical approach. Construction of the EDFs requires a certain number of samples to achieve a sufficient precision. However, this requirement may be violated in some atomic regions of the data. Therefore, the statistical tests obviously cannot be used in these regions. We display such regions with an additional color that is not associated with the

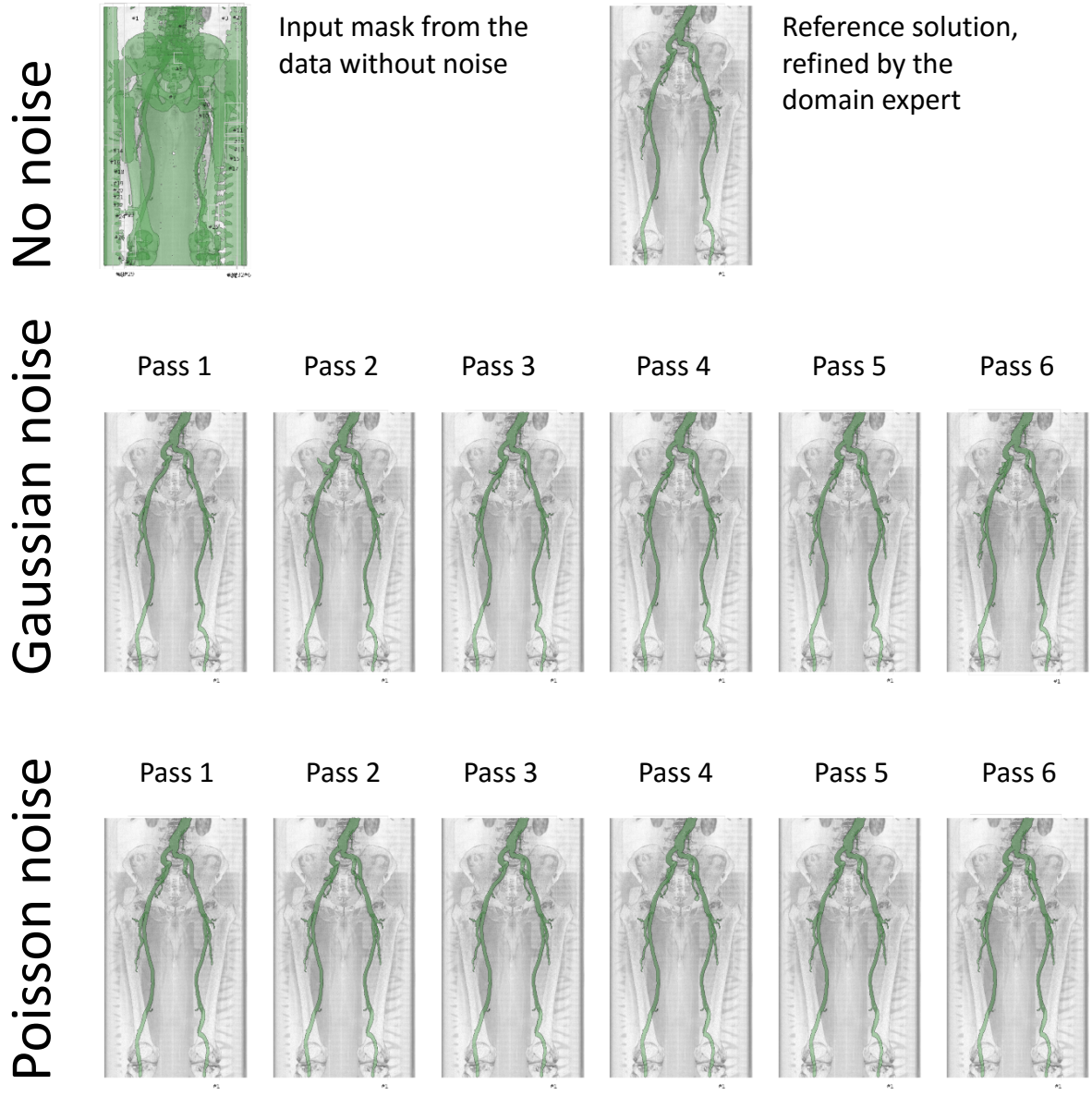

Figure 19: Comparison of the results with respect to different noise models.

Table 1: Detailed information on the volume editing experiment with artificial noise

	Gaussian noise					
	Pass 1	Pass 2	Pass 3	Pass 4	Pass 5	Pass 6
Standard deviation σ	1	2	4	8	16	32
Estimated PSNR (dB)	65.1	59.9	54.3	48.4	42.6	36.9
Editing time (sec.)	103	79	116	150	86	101
Number of steps	8	3	7	6	5	5
Quality of result	0.972	0.949	0.971	0.943	0.935	0.884
	Poisson noise					
	Pass 1	Pass 2	Pass 3	Pass 4	Pass 5	Pass 6
Mean μ	1	2	4	8	16	32
Estimated PSNR (dB)	66.3	63.3	60.4	57.3	54.3	51.4
Editing time (sec.)	91	81	116	122	192	102
Number of steps	4	3	7	10	9	5
Quality of result	0.979	0.980	0.971	0.974	0.969	0.948

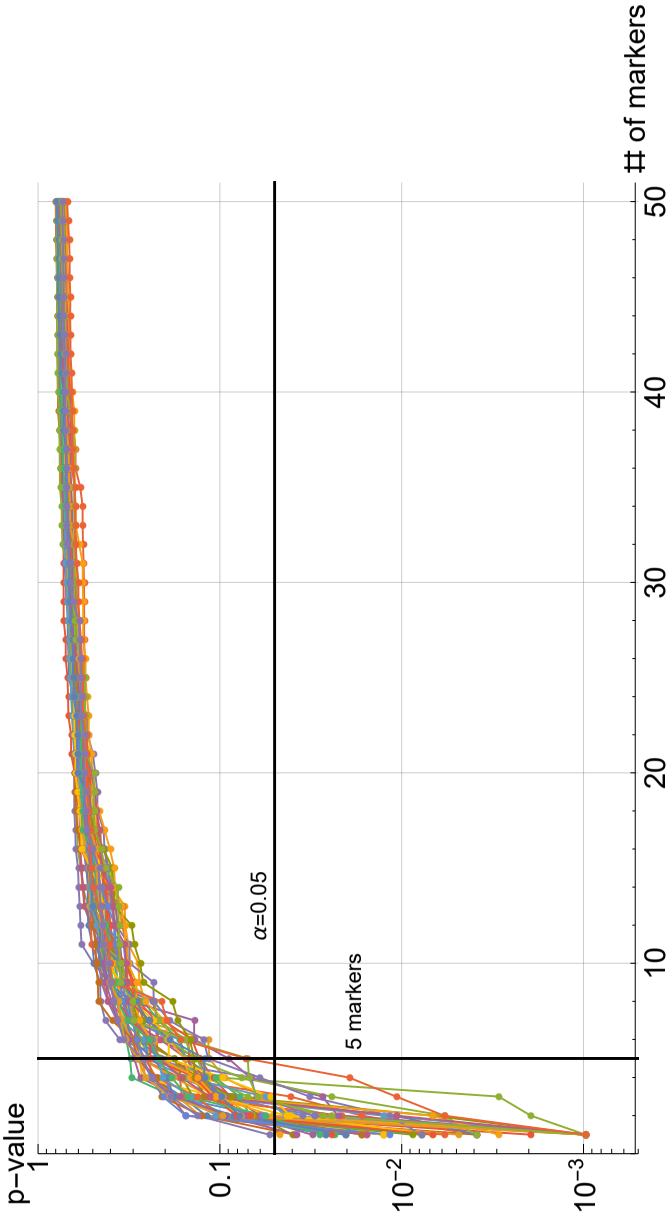

Figure 20: Median p-values for different number of markers, placed randomly inside the vessels. Each polyline represents one marker set. 50 marker sets were tried.

ranking color scheme. The user has to evaluate dissimilarities in this kind of regions, which is a conventional data-analysis task. For example, this happens with the few small vessels of the human lower extremities in CTA data. One possibility to address this issue is to improve the sampling of the data acquisition technique.

The proper atomic regions realization is crucial for our statistical approach. If the atomic regions for some reason do not reflect important features of the data, then the results of our method, namely the composite regions, their hierarchy, and the statistical information, may be sub-optimal for the data exploration task. If this happens, the user is required to check the entire hierarchy manually in order to find regions that satisfy the chosen objective. For example, in the case of the three-dimensional spatial data it is important to capture spatial features. Our specialized realization, described in Section 5.2, does this by representing the data with the influence zones. In certain cases, such a representation is sub-optimal, *e.g.*, if two different structures are connected with each other by a large contact area. As a result, these two structures are not separated at the level of the atomic regions, and, therefore, they are treated by our statistical approach as belonging together. Particularly, this situation occurs in regions, where the small vessels touch the bones in the human lower extremities (CTA data). Additional domain knowledge may rectify this deficiency.

If there were a systematic difference between distributions of conceptually same regions of the data, caused by the data acquisition modality, our statistical approach would over-estimate the dissimilarity significance. In particular, data from EM (Electron Microscopy) imaging has a significant variation between the slices, acquired from the specimen, which is physically cut into thin slabs. One way to deal with such an over-estimation is to pre-process the data, reducing the undesired variation in it.

In case of degenerate distributions, which have only a single outcome value, the statistical tests may report extreme p-values – zeros and ones. However, this does not pose a limitation for our statistical approach, as the comparison by the p-values is still valid. Our statistical method does not work with data, where regions that are different according to the domain-specific logic exhibit the same data-value distribution. However, in this case other general approaches of comparison would probably fail as well, as domain-specific knowledge is required to differentiate such regions, indistinguishable by the data values alone.

7 Conclusion

We proposed a novel abstract concept for statistically quantifying dissimilarities between arbitrary regions of m-dimensional data. The dissimilarity significances are computed by hypothesis testing, based on robust and sound statistical concepts. To facilitate data exploration, we represent the data with different levels of detail. At each level, we localize the regions with the most and the least significant dissimilarities, aiding the user during data exploration and analysis. We evaluated the generality of our method with two concrete applications: temporal data exploration and segmentation editing. Our proposed data exploration protocol streamlines the user interaction in both scenarios, which is strengthened by an evaluation with domain experts.

References

[1] J. M. Kniss, R. Van Uitert, A. Stephens, G. S. Li, T. Tasdizen, and C. Hansen, "Statistically quantitative volume visualization," in *Proceedings of IEEE Visualization Conference*, vol. 6, 2005, p. 37. DOI: 10.1109/VISUAL.2005.1532807.

[2] T. Tasdizen, S. P. Awate, R. T. Whitaker, and N. L. Foster, "MRI tissue classification with neighborhood statistics: A nonparametric, entropy-minimizing approach," *Lecture Notes in Computer Science*, vol. 3750, pp. 517–525, 2005. DOI: 10.1007/11566489_64.

[3] C. Lundström, P. Ljung, and A. Ynnerman, "Local histograms for design of transfer functions in direct volume rendering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 6, pp. 1570–1579, 2006. DOI: 10.1109/TVCG.2006.100.

[4] C. Heinzl, J. Kastner, T. Möller, and E. Gröller, "Statistical analysis of multi-material components using Dual Energy CT," in *Proceedings of Vision, Modeling, and Visualization*, 2008, p. 179.

[5] C. R. Johnson and J. Huang, "Distribution-driven visualization of volume data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 5, pp. 734–746, 2009. DOI: 10.1109/TVCG.2009.25.

[6] A. Saad, G. Hamarneh, and T. Möller, "Exploration and visualization of segmentation uncertainty using shape and appearance prior information," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1366–1375, 2010. DOI: 10.1109/TVCG.2010.152.

[7] M. Haidacher, D. Patel, S. Bruckner, A. Kanitsar, and M. E. Gröller, "Volume visualization based on statistical transfer-function spaces," in *Proceedings of PacificVis*, 2010, pp. 17–24. DOI: 10.1109/PACIFICVIS.2010.5429615.

[8] C. M. Jarque and A. K. Bera, "A test for normality of observations and regression residuals," *International Statistical Review*, vol. 55, no. 2, pp. 163–172, 1987. DOI: 10.2307/1403192.

[9] B. L. Welch, "The generalization of 'Student's' problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1, pp. 28–35, 1947. DOI: 10.2307/2332510.

[10] J. S. Praßni, T. Ropinski, J. Mensmann, and K. Hinrichs, "Shape-based transfer functions for volume visualization," in *Proceedings of PacificVis*, 2010, pp. 9–16. DOI: 10.1109/PACIFICVIS.2010.5429624.

[11] A. Karimov, G. Mistelbauer, T. Auzinger, and S. Bruckner, "Guided volume editing based on histogram dissimilarity," *Computer Graphics Forum*, vol. 34, pp. 91–100, 2015. DOI: 10.1111/cgf.12621.

[12] H. Hochheiser and B. Shneiderman, "Dynamic query tools for time series data sets: Timebox widgets for interactive exploration," *Information Visualization*, vol. 3, no. 1, pp. 1–18, 2004.

[13] P. Buono, A. Aris, C. Plaisant, A. Khella, and B. Shneiderman, "Interactive pattern search in time series," *Proceedings of SPIE*, vol. 5669, no. 1, pp. 175–186, 2005. DOI: 10.1117/12.587537.

[14] P. Buono, C. Plaisant, A. Simeone, A. Aris, B. Shneiderman, G. Shmueli, and W. Jank, "Similarity-based forecasting with simultaneous previews: A river plot interface for time series forecasting," in *Proceedings of International Conference on Information Visualisation*, 2007, pp. 191–196. DOI: 10.1109/IV.2007.101.

[15] M. Bögl, W. Aigner, P. Filzmoser, T. Lammarsch, S. Miksch, and A. Rind, "Visual analytics for model selection in time series analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2237–46, 2013. DOI: 10.1109/TVCG.2013.222.

[16] M. A. Stephens, "EDF statistics for Goodness-of-Fit and some comparisons," *Journal of the American Statistical Association*, vol. 69, no. 347, pp. 730–737, 1974. DOI: 10.1080/01621459.1974.10480196.

[17] T. Duong, B. Goud, and K. Schauer, "Closed-form density-based framework for automatic detection of cellular morphology changes," in *Proceedings of the National Academy of Sciences*, vol. 109, 2012, pp. 8382–8387. DOI: 10.1073/pnas.1117796109.

[18] A. N. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione," *Giornale dell'Istituto Italiano degli Attuari*, vol. 4, pp. 83–91, 1933.

[19] N. Smirnov, "On the estimation of the discrepancy between empirical curves of distribution for two independent samples," *Rec. Math. [Mat. Sbornik] N.S.*, vol. 6, no. 48, pp. 3–26, 1939.

[20] T. B. Arnold and J. W. Emerson, "Nonparametric Goodness-of-Fit tests for discrete null distributions," *The R Journal*, vol. 3, no. 2, pp. 34–39, 2011.

[21] J. M. Dufour and A. Farhat. (2001). Exact nonparametric two-sample homogeneity tests for possibly discrete distributions. Accessed: 2015-Oct-25, [Online]. Available: http://hdl.handle.net/1866/362.

[22] J. T. Praestgaard, "Permutation and bootstrap Kolmogorov-Smirnov tests for the equality of two distributions," *Scandinavian Journal of Statistics*, vol. 22, no. 3, pp. 305–322, 1995.

[23] B. F. J. Manly, *Randomization, Bootstrap, and Monte Carlo Methods in Biology*, Second. Chapman and Hall, London, 1997.

[24] D. A. Jackson and K. M. Somers, "Are probability estimates from the permutation model of Mantel's test stable?" *Canadian Journal of Zoology*, vol. 67, no. 3, pp. 766–769, 1989. DOI: 10.1139/z89-108.

[25] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Frannery, *Numerical Recipes in FORTRAN 77; The Art of Scientific Computing*, Second. The Press Syndicate of the University of Cambridge, 1993, ISBN: 052143064X. DOI: 10.1016/0378-4754(93)90043-T.

[26] T. W. Anderson, "On the distribution of the two-sample Cramér-von Mises criterion," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1148–1159, 1962. DOI: 10.1214/aoms/1177704477.

[27] A. N. Pettitt, "A two-sample Anderson-Darling rank statistic," *Biometrika*, vol. 63, no. 1, pp. 161–168, 1976. DOI: 10.2307/2335097.

[28] E. D. Feigelson and G. J. Babu. (2012). Beware the Kolmogorov-Smirnov test! Accessed: 2015-Oct-25, [Online]. Available: https://asaip.psu.edu/Articles/beware-the-kolmogorov-smirnov-test.

[29] T. Nichols, M. Brett, J. Andersson, T. Wager, and J. B. Poline, "Valid conjunction inference with the minimum statistic," *NeuroImage*, vol. 25, no. 3, pp. 653–660, 2005. DOI: 10.1016/j.neuroimage.2004.12.005.

[30] S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. J. Williams, *The American Soldier: Adjustment during Army Life*. Princeton University Press, Princeton, 1949, vol. 1.

[31] T. Liptak, "On the combination of independent tests," *Magyar Tud. Akad. Mat. Kutato Int. Kozl.*, vol. 3, pp. 171–197, 1958.

[32] F. Mosteller and R. R. Bush, "Selected quantitative techniques," in *Handbook of Social Psychology*, vol. 1, Addison-Wesley, Cambridge, 1954, pp. 289–334.

[33] M. C. Whitlock, "Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach," *Journal of Evolutionary Biology*, vol. 18, no. 5, pp. 1368–1373, 2005. DOI: 10.1111/j.1420-9101.2005.00917.x.

[34] A. M. Winkler, S. M. Smith, and T. E. Nichols, *Non-parametric combination for analyses of multimodal imaging*, Poster at the Organization for Human Brain Mapping in Seattle, 16-20 June, 2013.

[35] E. S. Edgington, "An additive method for combining probability values from independent experiments," *The Journal of Psychology*, vol. 80, pp. 31–363, 1972. DOI: 10.1080/00223980.1972.9924813.

[36] R. A. Fisher, *Statistical Methods for Research Workers*. Oliver and Boyd (Edinburgh), 1925, ISBN: 0-05-002170-2.

[37] K. J. Friston, A. P. Holmes, C. J. Price, C. Büchel, and K. J. Worsley, "Multisubject fMRI studies and conjunction analyses," *NeuroImage*, vol. 10, no. 4, pp. 385–396, 1999. DOI: 10.1006/nimg.1999.0484.

[38] L. H. C. Tippett, *The Methods of Statistics*, ser. Wiley publications in statistics. Williams and Norgate, 1952.

[39] R. D. Cousins, "Annotated bibliography of some papers on combining significances or p-values," pp. 1–15, 2008. arXiv: 0705.2209. [Online]. Available: http://arxiv.org/abs/0705.2209.

[40] K. Moreland, "Diverging color maps for scientific visualization," in *Proceedings of 5th International Symposium on Visual Computing*, 2009, pp. 92–103. DOI: 10.1007/978-3-642-10520-3_9.

[41] W. J. Cody, "Algorithm 715: SPECFUN - a portable FORTRAN package of special function routines and test drivers," *ACM Transactions on Mathematical Software*, vol. 19, no. 1, pp. 22–32, 1993. DOI: 10.1145/151271.151273.

[42] M. J. Wichura, "Algorithm AS 241: The percentage points of the normal distribution," *Applied Statistics*, vol. 37, no. 3, pp. 477–484, 1988. DOI: 10.2307/2347330.

[43] H. K. Hahn and H.-O. Peitgen, "IWT – interactive watershed transform: A hierarchical method for efficient interactive and automated segmentation of multidimensional grayscale images," in *Proceedings of SPIE Medical Imaging*, vol. 5032, 2003, pp. 643–653. DOI: 10.1117/12.481097.

[44] A. Karimov, G. Mistelbauer, J. Schmidt, P. Mindek, E. Schmidt, T. Sharipov, S. Bruckner, and E. Gröller, "ViviSection: Skeleton-based volume editing," *Computer Graphics Forum*, vol. 32, pp. 461–470, 2013. DOI: 10.1111/cgf.12133.

[45] T. C. Lee, R. L. Kashyap, and C. N. Chu, "Building skeleton models via 3D medial surface/axis thinning algorithms," *Lecture Notes in Computer Science*, vol. 56, pp. 462–478, 1994. DOI: 10.1006/cgip.1994.1042.

[46] D. Fleischmann, M. Straka, J. Roos, J. Lammer, R. Schernthaner, R. Schernthaner, M. Šrámek, A. Varchola, V. Solcany, and E. Gröller. (2007). AngioVis framework. Accessed: 2015-Oct-25, [Online]. Available: http://www.angiovis.org/.

[47] J. Brooke, "SUS: A "quick and dirty" usability scale," in *Usability Evaluation in Industry*, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and A. L. McClelland, Eds., London: Taylor and Francis, 1996, ch. 21, pp. 189–194.