Continuous Scatterplots

Sven Bachthaler and Daniel Weiskopf, *Member, IEEE Computer Society*

Abstract—Scatterplots are well established means of visualizing discrete data values with two data variables as a collection of discrete points. We aim at generalizing the concept of scatterplots to the visualization of spatially continuous input data by a continuous and dense plot. An example of a continuous input field is data defined on an n-D spatial grid with respective interpolation or reconstruction of in-between values. We propose a rigorous, accurate, and generic mathematical model of continuous scatterplots that considers an arbitrary density defined on an input field on an n-D domain and that maps this density to m-D scatterplots. Special cases are derived from this generic model and discussed in detail: scatterplots where the n-D spatial domain and the m-D data-attribute domain have identical dimension, 1-D scatterplots as a way to define continuous histograms, and 2-D scatterplots of data on 3-D spatial grids. We show how continuous histograms are related to traditional discrete histograms and to the histograms of isosurface statistics. Based on the mathematical model of continuous scatterplots, respective visualization algorithms are derived, in particular for 2-D scatterplots of data from 3-D tetrahedral grids. For several visualization tasks, we show the applicability of continuous scatterplots. Since continuous scatterplots do not only sample data at grid points but interpolate data values within cells, a dense and complete visualization of the data set is achieved that scales well with increasing data set size. Especially for irregular grids with varying cell size, improved results are obtained when compared to conventional scatterplots. Therefore, continuous scatterplots are a suitable extension of a statistics visualization technique to be applied to typical data from scientific computation.

Index Terms—Scatterplot, histogram, continuous frequency plot, interpolation.

1 Introduction

Scatterplots have been proven successful and useful diagramming techniques in descriptive statistics and information visualization. They take discrete data points with two data dimensions as input, and produce a 2-D plot of those data points by drawing respective dots on a diagram with two orthogonal axes representing the two data dimensions. Scatterplots are effective in displaying relationship in the data, such as correlation or other patterns.

In this paper, we utilize the undisputed visualization power of scatterplots and extend them to display continuous data typically generated in computational sciences. The method of this paper fits in the recent trend of combining statistical visualization methods like scatterplots with scientific visualization methods like volume or flow visualization.

We define the term *continuous scatterplot* as follows. First, input data is no longer a collection of discrete data points but a field of data values defined on a continuous domain. Typically, the domain is 2-D or 3-D and has intrinsic spatial embedding; the dimensionality of the domain increases by one if time dependency is included. Data representation often relies on a grid with respective interpolation or approximation schemes. Please note that the data field defined on the continuous domain is not necessarily continuous, i.e., even non-continuous functions can be visualized. The second aspect of a continuous scatterplot is that the output is continuous: instead of a collection of discrete points, a continuous density is drawn on the 2-D diagram, providing a continuous frequency description of two data dimensions.

Our model of continuous scatterplots and respective computations is generic in the sense that both the spatial domain of the input field and the dimension of the scatterplot can be chosen freely. In practice, the dimensionality of the input-field domain is determined by the data input, which mostly is restricted to dimensions 2–4. From a perceptual point of view, useful output dimensionalities are 1 (i.e., continuous histogram) and 2 (i.e., scatterplot in its original sense). Although there is previous work on 3-D discrete scatterplots, their perceptual effectiveness is unclear. To visualize higher-dimensional data, 2-D continuous scatterplots can be combined to respective scatterplot matrices.

The main contributions of this paper are: (i) We present a mathematical model for generic continuous scatterplots that maps an arbitrary density function from the n-D input domain to the m-D scatterplot domain in order to build the continuous frequency plot; here, interpolation and reconstruction schemes on the input domain are taken into account to construct a continuous frequency map. (ii) Generic continuous scatterplots are related to existing models like discrete histograms and scatterplots or continuous probability density functions to demonstrate that those concepts are generalized by continuous scatterplots. (iii) We propose an efficient computational model for the special, yet important case of 2-D continuous scatterplots of 3-D data defined on spatial tetrahedral grids.

We see the following benefits of our visualization approach: (i) The generic mathematical model of continuous scatterplots provides a solid and reliable basis for many variants of frequency plots of continuous data. (ii) Continuous density plots scale well with increasing data set size and resolution because they are not subject to overplotting issues, which for example occur due to dot plotting in discrete 2-D scatterplots. (iii) Continuous scatterplots are parameter-free. In contrast, the quality of discrete histograms depends on the choice of bucket size; a similar observation can be made for density constructions for discrete 2-D scatterplots. (iv) Since continuous scatterplots build a scalar density function, well known scalar-field visualization techniques can be applied for final plotting. For example, we use color mapping for 2-D scatterplots, and illustrate its usefulness by showing visualization results for a few typical visualization tasks. (v) Finally, continuous scatterplots readily fit in any visualization system or pipeline that supports discrete histograms or scatterplots. Therefore, continuous scatterplots may be employed widely in multi-attribute visualization of continuous data.

2 Related Work

Scatterplots, histograms, and similar plotting techniques are well-known and accepted diagramming methods in descriptive statistics and information graphics. An overview of statistics and respective visualization can be found in the books by Utts [15] and Chambers et al. [3], and the illustrated reference book by Harris [9]. In this paper, we view scatterplots as a way to visualize frequency distributions of 2-D data: scatterplots show frequency by the density of points drawn on the 2-D plot. Probably the most popular visualization of frequency distribution is based on histograms [10], designed for 1-D data.

Traditionally, statistics visualization is applied to discrete data

• *The authors are with VISUS (Visualization Research Center), Universität Stuttgart, Nobelstr. 15, 70569 Stuttgart, Germany, E-mail: {bachthaler, weiskopf}@visus.uni-stuttgart.de.*

Manuscript received 31 March 2008; accepted 1 August 2008; posted online 19 October 2008; mailed on 13 October 2008.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

points and, thus, continuous interpolation of data and continuous visualization has not been playing an important role. However, there is a new trend in the visualization community to apply techniques from statistics visualization to data fields defined on continuous domains (e.g., from areas like flow visualization or volume visualization) in order to explore multi-attribute fields. One example is the SimVis system [5], which employs multiple coordinated plots with brushing-and-linking in order to explore simulation data such as that from computational fluid dynamics; here, scatterplots are key visualization components. Another example is the specification of 2-D transfer functions [11] using 2-D plots that show the frequency of data entries of 2-D data points. While the combination of statistics visualization and continuous-domain fields is relatively new for scientific visualization, there is a related discipline that has been dealing with this issue for some time: geostatistics analyzes data typically given on 2-D spatial domains [12]. Although geostatistics employs some continuous interpolation techniques such as kriging, it mostly applies statistics to discretely sampled data. Similarly, the aforementioned combination of statistics visualization and scientific visualization relies on discrete data points (e.g., sampled grid points or vertices) and does not consider the continuous domain of the data field. In contrast, our approach takes into account the continuous spatial domain, the interpolation or reconstruction scheme, and the grid structure of the input data; and it also produces continuous versions of statistics plots.

Although descriptive statistics focuses on discrete data, there are a few examples of continuous plotting. One example is the extension of discrete histograms to continuous frequency distribution graphs: discrete columns of the histogram are replaced by a smooth curve that connects discrete points of the corresponding histogram [9]. While this approach leads to a continuous plot by means of interpolation within the frequency graph, it does not consider the continuous nature of the input data field but still processes discrete samples. The same holds for cumulative frequency graphs, also called ogives. Since the underlying model of such frequency graphs can be identified with probability density, the wealth of measure and probability theory [1] and mathematical analysis can be applied. Some aspects of integration theory will be used in the mathematical derivations and discussions of this paper.

Finally, the work by Carr et al. [2] is partly related to this paper. They investigate histograms for isosurface statistics and show that these histograms represent spatial function distributions. In Section 3.3, we demonstrate that the continuous histograms of this paper (for the special case of 3D scalar fields) share the same computation of isosurface area as in [2], but that we also take into account density transformations, which are important for generic data analysis targeted by our paper. Independently from us, Scheidegger et al. [13] have very recently revisited the histograms for isosurface statistics: using Federer's coarea formula [6], they take into account the density of isosurfaces and derive the same expression for isosurface histograms as in our paper.

3 Mathematical Model

This section presents the mathematical model of generic continuous scatterplots. We first introduce required terminology and definitions before we describe the generic model. Later parts of this section derive specific results for special cases like 2-D scatterplots or 1-D histograms. In this context, these continuous versions are compared with their well-known discrete counterparts.

3.1 Generic Model

The computation of continuous scatterplots needs two different domains: the n-dimensional domain on which the input field is defined, and the m-dimensional domain of the scatterplot onto which the input field is mapped. We denote the first one as *spatial domain* because it typically describes 2-D or 3-D spatial positions. Despite this terminology, the spatial domain may also include time dimension; in fact, any continuous field domain is supported. The second kind of domain is denoted as *data domain* because it represents the multi-attribute data

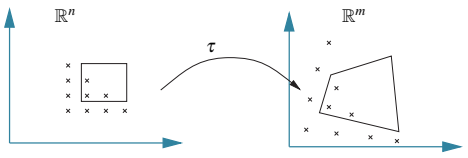

Fig. 1. Change of point density, by applying the mapping τ from the spatial domain of the input data (left) to the data domain of the scatterplot (right). Sampling density is indicated by dots; the two quadrilaterals visualize the mapping of an extended area. Here, both spatial and data domains have dimension 2.

values of the input field. Typical examples are $m = 1$, which corresponds to a histogram for one data component, or $m = 2$, which corresponds to a 2-D scatterplot. The input field to be visualized can be represented mathematically by a map from spatial domain to data domain: $\tau : \mathbb{R}^n \to \mathbb{R}^m$. The map τ can represent all typical scientific data, including 3-D scalar fields, 3-D vector fields, or multi-attribute fields.

The problem of constructing a continuous scatterplot can then be formulated as finding a density function σ defined on the data domain:

$$\sigma : \mathbb{R}^m \longrightarrow \mathbb{R}, \quad \xi \longmapsto \sigma(\xi)\,, \tag{1}$$

which represents continuous frequency and depends on τ. In fact, the mathematical basis of the continuous scatterplot is an operator that maps the function τ to the function σ.

To construct this operator, we derive a continuous description by starting from well-known discrete scatterplots and considering the limit process for infinitely dense data points. This approach is similar to deriving continuum mechanics from systems of discrete mass points (see, for example, [7, Ch. 12]). Figure 1 illustrates the discrete particle model for the example $n = m = 2$. The derivation is based on two assumptions: (i) points in the spatial domain are given according to some kind of density description (typically, uniform density), and (ii) the mapping τ does not change the number of points. The second assumption is identical to mass conservation if mass points are considered. The limit process for infinitely dense particles leads to a replacement of particle mass by mass density.

According to assumption (i), the mass (i.e., sampling) density s is known in the spatial domain, with $s : \mathbb{R}^n \to \mathbb{R}, x \mapsto s(x)$. The mass M of a covered volume $V \subset \mathbb{R}^n$ is $M = \int_V s(x)\mathrm{d}^n x$. Similarly, the density σ in the data domain can be integrated to compute the mass of the covered volume $\Phi \subset \mathbb{R}^m$ according to $\int_\Phi \sigma(\xi)\,\mathrm{d}^m\xi$. Note that our notation uses Latin characters for quantities related to the spatial domain and Greek characters for quantities related to the data domain; lower case letters denote scalar or vector values, uppercase letters denote volumes.

If V and Φ are related by $\tau(V) = \Phi$, mass conservation under the transformation τ implies

$$M = \int_V s(x)\,\mathrm{d}^n x = \int_{\Phi=\tau(V)} \sigma(\xi)\,\mathrm{d}^m\xi\,, \tag{2}$$

which determines the unknown density function σ for a given input density s because (2) has to hold for any volume V in the spatial domain. Rewriting (2) leads to the alternative formulation

$$\int_\Phi \sigma(\xi)\,\mathrm{d}^m\xi = \int_{\tau^{-1}(\Phi)} s(x)\,\mathrm{d}^n x\,, \tag{3}$$

which equally determines σ because (3) has to hold for any volume Φ in the data domain. The inverse map $\tau^{-1}(\Phi)$ is well defined even if τ is not invertible because here we work on maps of sets, not of single function values.

Note that σ is only indirectly defined in (2) or (3) via the effect of integration. For the generic case of scatterplot computation, this indirect definition of σ is required so that we can support not only regular functions but also distributions (generalized functions) like Dirac

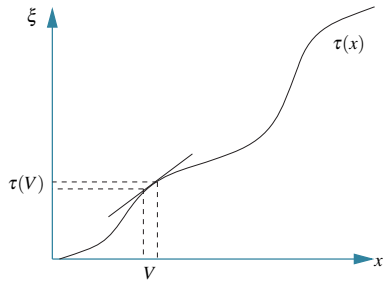

Fig. 2. Mapping of intervals from the spatial domain x to the data domain ξ. The ratio of the original and mapped intervals, which is given by the reciprocal of the slope of τ (slope is indicated by the straight line), is a measure for the change of density from the spatial to the data domain.

δ distributions. We refer to the textbook by Griffel [8] for an introduction to distributions and functional analysis. In fact, the earlier definition of σ as a regular function in (1) is extended to allow for distributions as well. We will show later that δ distributions are useful to build a relationship between continuous and discrete scatterplots, and to allow for scatterplots and histograms of (partly) constant functions. We will also show that the indirect formulation of (2) and (3) can be rewritten to directly compute σ for some special, yet important cases.

3.2 Case $m = n$

Let us now consider the special case of equal dimension $m = n$ in order to compute σ. Since σ is directly based on τ, we split this discussion into several parts that cover different possible subcases depending on the properties of τ. First, we will assume that τ is differentiable. Here, we will consider both subcases, where τ is a diffeomorphism or no diffeomorphism. Later, we will extend our discussion to non-differentiable τ.

Assuming that τ is a diffeomorphism from $\mathbb{R}^n \to \mathbb{R}^{m=n}$, the integration variable ξ can be substituted by x in (2) according to the transformation theorem for integrals:

$$\int_{\tau(V)} \sigma(\xi)\,\mathrm{d}^{m=n}\xi = \int_V \sigma(\tau(x))\,|\det(D\tau)(x)|\,\mathrm{d}^n x \stackrel{!}{=} \int_V s(x)\,\mathrm{d}^n x\,, \quad (4)$$

where $D\tau$ denotes the derivative of the map τ, i.e., the $n \times n$ Jacobi matrix. Note that the determinant is a volume measure, representing the volume spanned by the partial derivatives of τ. Since the second equality of (4) has to hold for any $V \subset \mathbb{R}^n$, the integrands need to be equal, which leads to

$$\sigma(\xi) = \frac{s(\tau^{-1}(\xi))}{|\det(D\tau)(\tau^{-1}(\xi))|}\,. \quad (5)$$

Figure 2 illustrates the density mapping according to (5) for a 1-D example $m = n = 1$. Geometrically, the density ratio σ/s in a small neighborhood is given by the ratio of the covered lengths, $V/\tau(V)$, which in turn is computed by the reciprocal of the slope of τ. Please note that a similar approach is followed for the inversion of cumulative distribution functions in order to map probability distributions or derive histogram equalization.

When $\det(D\tau) = 0$, then τ is not a diffeomorphism and (5) cannot be used to define σ. Such a case can occur, e.g., when τ contains regions of constant value, extremal points, or other points with vanishing $\det(D\tau)$. Let us first consider the case of constant value. Figure 3 illustrates the example of a piecewise constant function for dimensionality $m = n = 1$. In the spatial domain, a piecewise constant function can be modeled as $\tau(x) = \sum_i \tau_i \chi_i(x)$, with the characteristic function

$$\chi_i(x) = \begin{cases} 1 & \text{if } x \in \text{Cell}(i) \\ 0 & \text{else} \end{cases}\,.$$

It is assumed that the spatial domain is partitioned into cells i. Then, τ has constant value τ_i within cell i. Assuming constant density in the spatial domain, $s = 1$, the density in the data domain is:

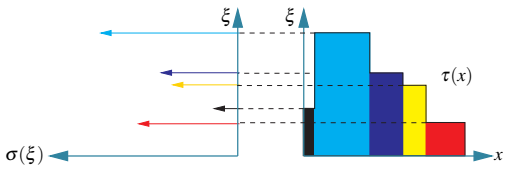

Fig. 3. Continuous histogram (left) of a piecewise constant function (right). The histogram is rotated by 90 degrees to visualize data correspondence between the function plot and the histogram. δ peaks are indicated by arrows; the length of a δ peak corresponds to the width of the respective interval in the function plot.

$$\sigma(\xi) = \sum_i \delta(\xi - \tau_i)\,\text{Size}(\text{Cell}(i))\,, \quad (6)$$

with Dirac delta δ[1]. The correctness of (6) can be verified by plugging (6) into the mass-conservation equation (3):

$$\int_\Phi \sigma(\xi)\,\mathrm{d}^n\xi = \sum_i \int_\Phi \delta(\xi - \tau_i)\,\text{Size}(\text{Cell}(i))\,\mathrm{d}^n\xi$$
$$= \sum_{\text{for all } i \text{ where } \tau_i \in \Phi} \text{Size}(\text{Cell}(i)) = \int_{\tau^{-1}(\Phi)} 1\,\mathrm{d}^n x\,.$$

Since $s = 1$, (3) is correctly met. Note that we have chosen $s = 1$ to simplify notation; an analogous proof would work for any choice of s, assuming that (6) includes s.

A different approach is taken for the other problematic case where $\det(D\tau) = 0$ at isolated parts. To be more precise, in this case, the determinant of the derivative vanishes at isolated null-sets (null-sets with respect to integration in the spatial domain). Here, we follow a two-step approach. First, regions where $\det(D\tau)(x) = 0$ are identified. These regions are denoted $\Gamma = \{x \in \mathbb{R}^n | \det(D\tau)(x) = 0\}$. Second, the null-set Γ and its image under the map τ are removed from the computation of density in (5), i.e., σ is not defined at those locations. Since integration over null-sets always yields 0, we can remove those isolated locations without affecting the conservation-of-mass model. The same approach is taken if the underlying data set is not continuous, i.e., τ is not differentiable (e.g., between cells). This results again in null-sets, which can be removed from the computation of the density. Finally, in theory, τ might be non-diffeomorphic on non-null-sets. However, we do not consider this case, since any realistic data set will meet the requirement that τ is non-differentiable or has vanishing $\det(D\tau)$ at the most at isolated null-sets. In particular, any grid-based data set with piecewise cell-oriented interpolation (which is most common in scientific visualization), meets this requirement.

To summarize the special case of $m = n$, we can compute σ by the following schematic algorithm. (i) For an extended volume (i.e., not a null-set) where τ is constant, a δ peak is associated that is weighted by the size of the volume. (ii) Isolated null-sets of non-diffeomorphism are identified and used to partition V into volumes where $\det(D\tau) \neq 0$. The null-sets themselves are removed from the computations, and intermediate σ are computed for each element of the partition. (iii) The intermediate results from (i) and (ii) are added. Step (iii) is valid because the defining equation (3) is linear. Figure 4 illustrates the schematic algorithm for an example of a continuous histogram with $m = n = 1$ and with constant density s in the spatial domain.

Our construction of continuous histograms allows us to reproduce discrete histograms. Please note that discrete input data has no attached spatial domain in traditional statistical analysis. We identify discrete data points with arbitrarily positioned cells in the spatial domain; all cells are chosen with equal size. Then, the continuous histogram consists of δ peaks at the data point values (see construction in

[1]The Dirac distribution has the property $\int \delta(x - x_0)\,\mathrm{d}^n x = 1$ if x_0 is in the integration domain

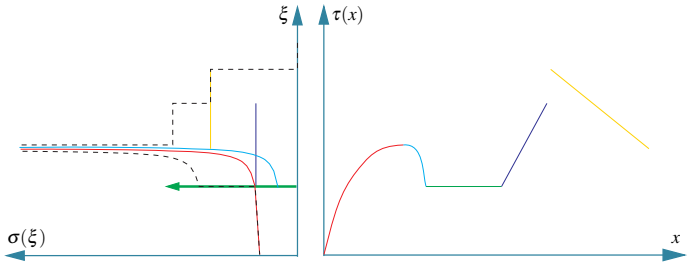

Fig. 4. General example of a continuous histogram (left) for the function τ (right). The spatial domain of τ is partitioned into intervals with non-vanishing derivative; intervals are color-coded (right image). Respective densities σ are shown in the left image (same color coding). The black dashed curve shows the sum of the intermediate densities σ, which should be overlaid with the δ peak (green) corresponding to the constant part of τ.

(6)), i.e., δ peaks in the continuous histogram correspond to entries in the discrete histogram. Typical bucketing of discrete histograms can be achieved by integrating the continuous histogram with box functions that represent the buckets.

3.3 Case $m < n$

Another common case is $m < n$, which arises for example for a 2-D scatterplot that shows two scalar attributes of a 3-D volume data set. Here, the dimension of the spatial domain is reduced when τ maps to the data domain. Therefore, the transformation theorem for integration, which was used in the previous subsection, does not apply. In particular, $\det(D\tau)$ does not exist, which means that we cannot use the same approach as in (4).

Figure 5 illustrates the geometry of the underlying problem for the case $n = 2$ and $m = 1$: a single point in the data domain corresponds to an infinite set of points in the spatial domain. In this example, the infinite set is the isoline within the spatial domain that corresponds to the isovalue in the data domain. This imbalance in dimensionality does no longer permit the transformation of differentials as in the previous subsection.

As in the generic discussion of Section 3.1, we consider two related volumes in the data and spatial domains: $\Phi \subset \mathbb{R}^m$ and $V = \tau^{-1}(\Phi) \subset \mathbb{R}^n$. To be more specific, let us restrict ourselves to a small neighborhood Φ around a point $\xi_0 \in \mathbb{R}^m$. To overcome the dimensionality problem, we split V into two parts: (i) the inverse image of the point ξ_0 (i.e., a generalized isocontour), and (ii) the perpendicular space around the inverse image of ξ_0. We denote $\tau^{-1}_{\text{normal}}(p)$ as the space that is normal to $\tau^{-1}(\xi_0)$ at a point $p \in \tau^{-1}(\xi_0)$ and that is also contained within $\tau^{-1}(\Phi)$. Here, we assume that τ is a smooth non-constant function, and therefore, the isocontour is smooth as well and the normal space is well defined. If τ is not smooth, then the spatial domain is split into piecewise smooth regions and the method is applied in a piecewise manner (our approach does not apply to completely non-smooth functions). If τ is (partly) constant, such a constant volume is separated out of the computation similar to the discussion in Section 3.2, leading to δ contributions.

In the regular case, by construction, the normal space has the same dimension m as the data domain, whereas the isocontour has dimension $(n-m)$. Figure 5 illustrates the isocontour and the normal space. Now, the integration in (3) can be split into isocontour and normal parts, similar to the approach of coarea computation [6]:

$$\int_{\Phi} \sigma(\xi)\,\mathrm{d}^m\xi = \int_{\tau^{-1}(\Phi)} s(x)\,\mathrm{d}^n x = \int_{\tau^{-1}(\xi_0)} \left(\int_{\tau^{-1}_{\text{normal}(\hat{x})}} s(\tilde{x})\,\mathrm{d}^m\tilde{x} \right) \mathrm{d}^{(n-m)}\hat{x}\,. \tag{7}$$

Within the normal space, we can use the same approach as in Section 3.2 because dimensionalities coincide and a computation based on derivatives is possible. Denoting the intermediate contribution to the density at points $\hat{x}$ by $\tilde{\sigma}_{\hat{x}}$, we re-label the inner integration in the righthand side of (7) according to:

$$\int_{\tau^{-1}_{\text{normal}(\hat{x})}} s(\tilde{x})\,\mathrm{d}^m\tilde{x} \overset{!}{=} \int_{\Phi} \tilde{\sigma}_{\hat{x}}(\tilde{\xi})\,\mathrm{d}^m\tilde{\xi}\,.$$

Assuming a diffeomorphism in the normal space, (5) can be adopted to obtain:

$$\tilde{\sigma}_{\hat{x}}(\tilde{\xi}) = \frac{s(\hat{x})}{|\mathrm{Vol}(D\tau)(\hat{x})|}\,,$$

where $\hat{x}$ and $\tilde{\xi}$ are related by $\tau(\hat{x}) = \tilde{\xi}$. Here, $|\mathrm{Vol}(D\tau)|$ replaces the determinant $|\det(D\tau)|$ in (5). The volume measure $|\mathrm{Vol}(D\tau)|$ is defined as the volume spanned by the partial derivatives of τ restricted to variations of parameters in the normal space $\tau^{-1}_{\text{normal}(\hat{x})}$. Figure 5 (left) illustrates the reciprocal volume measure $1/|\mathrm{Vol}(D\tau)|$ within the spatial domain. The volume measure is explicitly computed in Section 3.4 for the special case $m = 2$, $n = 3$, and below in this subsection for the case $m = 1$, $n = 3$.

For the final overall density, $\tilde{\sigma}_{\hat{x}}$ is integrated along the complete isocontour:

$$\sigma(\xi_0) = \int_{\tau^{-1}(\xi_0)} \frac{s(\hat{x})}{|\mathrm{Vol}(D\tau)(\hat{x})|}\,\mathrm{d}^{(n-m)}\hat{x}\,, \tag{8}$$

which completes the generic discussion of the case $m < n$.

Here, we include a comparison with the work on isosurface statistics by Carr et al. [2] because we can produce similar histograms by using $m = 1$ and $n = 3$. Their paper focuses on analyzing isosurface behavior, whereas our continuous scatterplots target visual data analysis in the full domain. This is the reason for different definitions of histograms. Carr et al. define their histogram as the volume of the inverse image, i.e., the area of the respective isosurface. With our notation, their histogram would read $\sigma(\xi_0) = \int_{\tau^{-1}(\xi_0)} 1\,\mathrm{d}^{(n-m)}\hat{x}$ instead of our computation in (8). Both approaches use the size of the isosurface (here, via integration over $\tau^{-1}(\xi_0)$). One (minor) difference is that we support a space-variant input density s. The major difference, however, is that we take into account $1/|\mathrm{Vol}(D\tau)(\hat{x})|$, whereas Carr et al. do not. For $m = 1$ and $n = 3$, $|\mathrm{Vol}(D\tau)(\hat{x})|$ is the magnitude of the gradient at $\hat{x}$. Put differently, we consider the neighborhood of values in the data domain and how they are affected by derivatives of τ, whereas Carr et al. use a point mapping from the data domain to the spatial domain. In this sense, our approach is related to, but no identical with, Legendre transformations that take into account derivatives (see the geometric interpretation of the Legendre transformation in [4, pp. 32–39] and its use for Hamiltonian and Hamilton-Jacobi mechanics described in [7]). Therefore, our definition of continuous scatterplots takes into account the behavior of τ in its full neighborhood; only in this way, it is possible to represent the transformation of sampling density. Scheidegger et al. [13] have independently derived the same weighting factor of $1/||\nabla\tau(\hat{x})||$ when revisiting Carr et al. [2]; i.e., for the case of isosurface histograms, both approaches lead to the same result.

A formal, mathematical advantage of our model in (2) and (3) is its generic applicability to any dimension of the spatial and data domains.

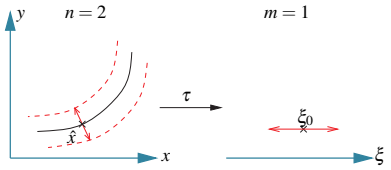

Fig. 5. Projection issue for the case $m < n$. A small interval (red) in the data domain around ξ_0 (right image) corresponds to a bent strip (outlined in red) around the isocontour (solid center line) with isovalue ξ_0 in the spatial domain (left image). The bent strip is the union of the normal spaces for all points on the isocontour. The red arrows in the left image mark the normal space at point $\hat{x}$.

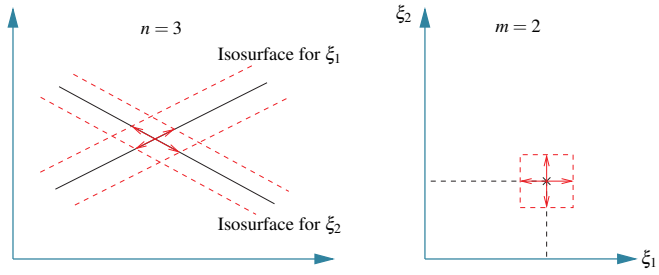

Fig. 6. In data domain, two values ξ_1 and ξ_2 are chosen (shown on the right side). These values correspond to isosurfaces represented by two solid lines in the spatial domain shown on the left side. The areas denoted by the small red arrows in the two domains correspond to each other. The reciprocal volume $1/|\mathrm{Vol}(D\tau)|$ is the quadrilateral that results from the intersection of the two stippled stripes around the isosurfaces.

In particular, densities s and σ are automatically adapted to the respective integration dimensions n and m—if actual physical units of mass density were used, they would have SI units (Système International d'unités) [kg/m^n] and [kg/m^m], respectively. Moreover, even (partly or completely) constant functions τ are supported in our model; in this case, δ distributions occur in σ. In contrast, the formal derivation of (5) by Carr et al. [2] is problematic because the integration measure in their expression $\int_{f^{-1}(h)} 1\,\mathrm{d}x$ is not explicitly specified but, from the context of their equation (3), should be d-dimensional. Typically, $f^{-1}(h)$ is an isosurface, which is a null-set, and therefore the integral vanishes. Alternatively, the integration measure could be adapted to the dimensionality of $f^{-1}(h)$. Then, their integration (5) would be fine as long as the dimensionality of $f^{-1}(h)$ is constant, i.e., problems would occur when the function is partly constant, leading to a mix of 2-D isosurfaces and 3-D isovolumes. Scheidegger et al. [13] resolve these problems by restricting the computation of isosurface histograms to 3-D scalar fields with non-vanishing gradient, implying integration on 2-D isosurfaces. The above issues with continuous distributions demonstrate the usefulness of our density-based definition of generic continuous scatterplots.

3.4 Case $m = 2$ and $n = 3$

This subsection addresses a special case of the above subsection: $m = 2$ and $n = 3$. This case is important in practical applications because typical data is given on a 3-D spatial domain and analyzed by 2-D scatterplots. The other important practical application is the computation of 1-D histograms for data on 3-D spatial domains; this application was covered at the end of the previous subsection.

For simplicity of discussion, we assume that τ is a smooth non-constant function so that for any choice of coordinates in the data domain, $\xi = (\xi_1, \xi_2)$, two smooth isosurfaces corresponding to ξ_1 and ξ_2 are obtained (for (partly) constant or non-smooth τ, a special treatment similar to Section 3.2 is required). Furthermore, we assume that τ is not degenerate so that the intersection of the two isosurfaces yields 1-D curves. Figure 6 illustrates the geometry of the scenario. Here, a zoomed-in view is shown; therefore, the smooth isosurfaces appear planar. In this case, (8) reads

$$\sigma(\xi_0) = \int_{\tau^{-1}((\xi_1,\xi_2))} \frac{s(\hat{x})}{|\mathrm{Vol}(D\tau)(\hat{x})|}\,\mathrm{d}\hat{x}\,, \tag{9}$$

where the integration is along the 1-D intersection of the two isosurfaces. The 2-D area $|\mathrm{Vol}(D\tau)|$ in (9) is spanned by the gradients $\partial\xi_1/\partial x$ and $\partial\xi_2/\partial x$. Figure 6 illustrates the respective reciprocal volume $1/|\mathrm{Vol}(D\tau)|$ carved out around the two isosurfaces. By using vector computations, the volume measure is computed as the cross product of the two gradients:

$$|\mathrm{Vol}(D\tau)| = \left|\left| \frac{\partial\xi_1}{\partial x} \times \frac{\partial\xi_2}{\partial x} \right|\right|. \tag{10}$$

In summary, the density σ is obtained by integration along the 1-D intersection curves of the two isosurfaces, weighted by the reciprocal of the magnitude of the cross product of the two respective gradients.

3.5 Case $m > n$

This subsection briefly discusses the remaining uncovered case $m > n$ in order to complete the description of cases. From a visualization point of view, this case is not very useful because the dimensionality of the scatterplot is higher than the dimensionality of the spatial domain, which adds visual complexity instead of reducing it.

The map τ from the spatial domain $\mathbb{R}^n$ to the data domain $\mathbb{R}^m$ leads to a coverage of the data domain by an n-D subset. For example, a 1-D spatial data set would typically result in a 1-D curve within a 2-D scatterplot, i.e., the support for the density σ would be that curve. The density distribution σ can be computed by applying the mapping from Section 3.2 within the support of σ, considering this support as an n-D manifold, and by allowing for δ distributions to obtain finite values when integrating over null-sets in the data domain.

4 Scatterplot Algorithm for Tetrahedral Meshes

In practice, the most important examples of continuous scatterplots are the cases $n = 3, m = 1$ (i.e., continuous histogram) and $n = 3, m = 2$ (continuous 2-D scatterplot). Both cases work on a 3-D spatial domain, which is common for scientific data. Even for time-dependent 4-D data, visualization is often restricted to showing 3-D time slices. The case of continuous histograms can be implemented similar to Carr et al. [2]; the only difference is the additional weighting by the reciprocal of the gradient magnitude and by the original density s. The extension of Scheidegger et al. [13] already includes the weighting by the reciprocal of the gradient magnitude and, thus, their implementation could be directly adopted. Therefore, this section focuses on the other case—the construction of continuous 2-D scatterplots.

According to (9) and (10), the intersection curve of two isosurfaces as well as the two gradients along the intersection curve need to be determined and combined by integration along the curve. The result of this computation depends on the functional behavior of the data field τ. Typically, volumetric data is given on a grid, and τ is reconstructed by piecewise cell-based interpolation within the grid. We focus on tetrahedral grids because they are naturally equipped with linear (barycentric) interpolation.

Since the overall density σ is based on the linear model of (2) and (3), we can construct σ by linear superposition of the contributions from tetrahedral cells. Therefore, the remaining question is how to compute σ for a single tetrahedron. Here, the linearity of barycentric interpolation simplifies the computation substantially because of the following reasons. First, isosurfaces within a tetrahedron are planes. Therefore, the intersection between two isosurfaces is a straight line (in the non-degenerate case). Second, the gradient within the cell is constant. Thus, the volume measure (10) is constant as well. In conclusion, σ is obtained by computing the length of the intersection of the two isosurfaces and dividing that value by the constant volume measure. Here, we assume a constant density s in the spatial domain.

To compute the isosurface intersection, we adopt the projected-tetrahedra algorithm by Shirley and Tuchman [14], designed for volume rendering of scalar fields on tetrahedral grids. The original algorithm projects tetrahedra onto the image plane, which is located within the spatial domain, whereas 2-D scatterplots need to project the tetrahedra to the data domain. This projection is achieved by interpreting (ξ_1, ξ_2) as coordinates for orthographic projection. The Shirley-Tuchman algorithm partitions the image footprint of the tetrahedron into a collection of a few triangles (up to four triangles), depending on the viewing direction. Within each triangle, parameters are interpolated linearly. The same kind of triangle partitioning is used for scatterplots. Here, the linearly interpolated parameter is the geometric depth of the tetrahedron in the spatial domain, computed at the corresponding data values (ξ_1, ξ_2). Figure 7 illustrates the computation of depth. Linear interpolation of depth within the triangle is correct because the underlying 3-D geometry is linear as well. The final σ value is obtained by dividing depth by the volume measure from (10).

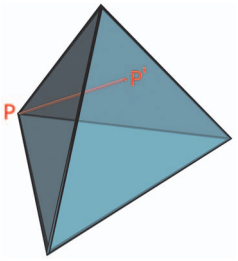

Fig. 7. This illustration shows how the distance is measured that is used to determine the density. At point P, we have to compute a depth value that corresponds to the density. This is done by calculating the distance between P and P' in the spatial domain of the tetrahedron. The point P' is the intersection point between the face opposite to P and the (ξ_1, ξ_2) isoline through P. (Please note that P' does not coincide with a vertex, except for degenerated cases).

The algorithm consists of the following steps:

1) Classify the tetrahedron based on its silhouette in the data domain. This step yields up to four triangles.
2) Attach data values (ξ_1, ξ_2) as 2-D geometric coordinates to the triangle vertices.
3) Determine the volume measure according to (10).
4) Compute the Euclidean distance between frontface and backface of the tetrahedron at the vertices. Attach this distance divided by the volume measure as texture coordinate to the vertices.
5) Render triangles. The volume-weighted distance is interpolated during scanline conversion and yields the density at the current fragment. Output the result to the framebuffer.

If a data value is constant within the tetrahedron, the corresponding σ is no longer a regular function, but a δ distribution. We approximate δ distributions by assigning a very large value (a constant defined within the implementation). In this way, even degenerate cases can be handled.

The overall density σ is obtained by rendering all tetrahedra with additive blending. Since additive blending is commutative, triangle sorting is not necessary prior to rendering. Finally, we apply color mapping by using a color lookup table in order to generate the final result.

The above algorithm works for any tetrahedral grid. For non-simplicial grids, we decompose cells into tetrahedra before rendering. In the common case of hexahedral cells, a decomposition in five tetrahedra per cell is employed. Since any 3-D grid can be approximated by triangulation, we can process any grid-based data set.

The projected-tetrahedra algorithm lends itself to acceleration by graphics hardware because rasterization of triangles and blending are efficiently supported by graphics hardware. Similarly, our 2-D scatterplot algorithm can be implemented using graphics hardware. Our implementation is based on C++ and DirectX. We have tested the implementation on a Windows PC with NVIDIA GeForce 8800 GTX GPU (768 MB). Steps 1–4 of the above algorithm are performed on the CPU, similar to traditional implementations of the Shirley-Tuchman algorithm. The results from all tetrahedra are combined by additive blending within a render-target texture. For appropriate blending quality, a framebuffer/texture format with 32-bit floating-point resolution is chosen. The render-target texture is used as input to another render process that applies a color table (implemented as 1-D dependent texture) to σ to generate the final image.

5 Results

In this section, we provide examples of both discrete and continuous scatterplots for three different visualization examples. Additional material can be found on our project web page.[2] Please note that constant input density $s = 1$ is used for all examples. The scatterplot functionality is part of a multi-attribute visualization system that also supports multiple coordinated views, brushing-and-linking, and volume visualization. For continuous scatterplots, the implementation described in the previous section was used. For discrete scatterplots, a similar GPU-based implementation was applied; here, points of the scatterplot are rendered via point sprites.

The first example in Fig. 8 shows scatterplots that support the user-guided specification of 2-D transfer functions for volume rendering [11]. The scatterplots visualize the "blunt-fin" data set, which is given on an unstructured grid derived from a curvilinear grid of resolution $40 \times 32 \times 32$. The scatterplot axes represent the scalar value and the gradient magnitude of the scalar field, respectively. Both scatterplots use a logarithmic color table to encode density values; low density is mapped to black/dark blue, mid-density values are shown in red, and high density values are yellow/white. We use this color table for density-to-color mapping for all result images of this section. As discussed by Kniss et al. [11], material boundaries lead to pronounced arc-like structures that can be selected by the user to highlight corresponding regions of the 3-D scalar field in the volume visualization within the spatial domain. Although both types of scatterplots show arc-like patterns, differences between discrete and continuous scatterplots are clearly visible. Unlike the discrete scatterplot, the continuous version provides a dense visualization that allows to spot interesting features more easily than in the discrete representation. The discrete scatterplot just uses the data at the grid points and ignores the underlying grid structure, whereas the continuous scatterplot takes into account the varying size and shape of grid cells by computing gradients within cells. Therefore, differences between discrete and continuous scatterplots may be particularly pronounced for unstructured or curvilinear grids compared to uniform grids with their constant cell size.

The next example in Fig. 9 shows the "tornado" data set. This data set is commonly used in flow visualization as a benchmark data set. It represents the 3-D velocity field of air flow of a simplified tornado. Data is given on a uniform grid of resolution 128^3. The two resulting scatterplot variants are compared in the upper part of Fig. 9. For this data set, we have several data dimensions that can be visualized in the scatterplot. In Fig. 9, we map the magnitude of the velocity to the horizontal axis and the velocity in z-direction to the vertical axis. In this way, different features of the "tornado" can be extracted, e.g., the inner part of the vortex region (Fig. 9b) or the outer boundary of the vortical structure (Fig. 9c). For the "tornado" data set, we also demonstrate brushing-and-linking. During this process, features in the scatterplot are identified and selected using a selection rectangle. In the volume visualization, voxels are highlighted if they correspond to the selected area defined in the scatterplot. Two different selections were made and their results are shown in the lower part of Fig. 9.

The third example is the IEEE Visualization 2004 Contest data set "Hurricane Isabel", depicted in Fig. 10. We show this data set in two different resolutions—a downsampled version with a size of $128 \times 128 \times 30$ and the original data set with a size of $500 \times 500 \times 100$. Air temperature is mapped to the horizontal axis, whereas air pressure is mapped to the vertical axis. Here, we demonstrate that continuous scatterplots are structurally independent of the resolution of the data set. In particular, the discrete scatterplot of the low-resolution data set induces misleading structures (i.e., the slanted, nearly vertically aligned clusters of points), which are not part of the data but due to the low sampling resolution. Moreover, while the visual result of a discrete scatterplot depends on the size of the individual points, continuous scatterplots are parameter-free.

For all three examples, continuous scatterplots show better visual quality than discrete scatterplots. In particular, discrete scatterplots tend to miss visual information in plots; those visual gaps require extra mental work by the user in order to close those gaps. In addition, some features are glossed over or are completely missing in discrete scatterplots. In contrast, continuous scatterplots provide guaranteed coverage of the relevant parts of the scatterplot domain and, thus, cannot miss important features. Furthermore, continuous scatterplots nicely fit in visualization systems with brushing-and-linking and multiple coordi-

[2] http://www.vis.uni-stuttgart.de/scatterplot

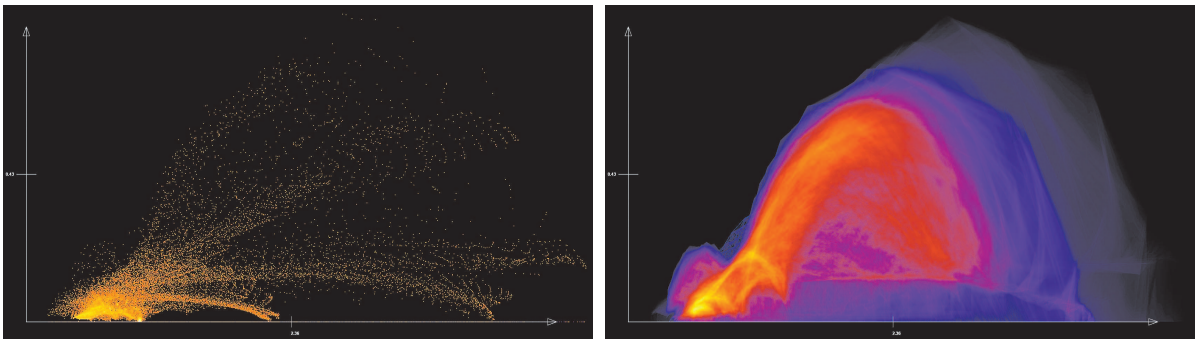

Fig. 8. The left side shows the discrete scatterplot of the "blunt-fin" data set, whereas the continuous version is shown on the right side. Both types of scatterplots visualize the scalar data value along the horizontal axis and the magnitude of the gradient along the vertical axis. Choosing these data dimensions, material and boundary identification is possible by finding arc-like structures.

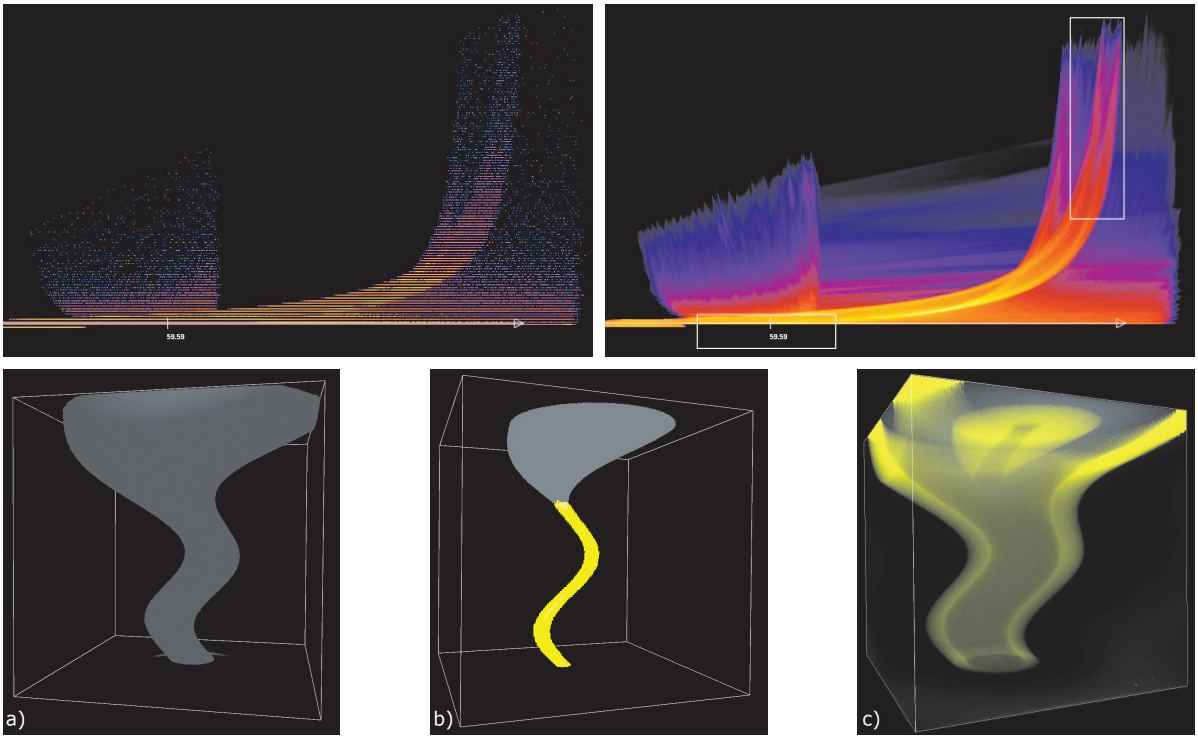

Fig. 9. In the upper part, both types of scatterplots are shown for the "tornado" data set. The upper-left image shows the discrete scatterplot, to the right is the continuous version. Both scatterplots employ the same color lookup table as the previous example. The lower part shows three volume visualizations of the data set. The lower-left image (a) shows the tornado visualized by a representative isosurface of velocity magnitude. The image in the middle (b) shows highlighted voxels (yellow) that were marked in the continuous scatterplot. This highlighting corresponds to the upper-right selection rectangle in the continuous scatterplot. The other selection rectangle in the lower-mid part of the continuous scatterplot highlights different voxels, as shown in the lower-right volume-visualization image (c). In image (c), highlighted voxels (yellow) and the velocity magnitude are simultaneously visualized by rather transparent volume rendering in order to show selected features at different depths. Therefore, we can see that different voxels than in (b) are highlighted, especially not the ones in the center of the tornado.

nated views.

6 Conclusion and Future Work

We have presented continuous scatterplots as a generalization of conventional scatterplots. One aspect of generalization is the support of any dimension of the domain of the data set and of the scatterplot. The other aspect of generalization is the extension to data defined on continuous domains. The basis for continuous scatterplots is provided in the form of a generic mathematical model. This mathematical model maps an arbitrary density value defined on an n-D input data set to m-D scatterplots. We have shown how continuous scatterplots are related to conventional discrete histograms and to histograms of isosurface statistics. In particular, the 2-D version of continuous scatterplots is, by construction, identical to conventional discrete scatterplots in the limit process of infinitely dense sample points. Therefore, continuous scatterplots lead to the same basic visual mapping as traditional histograms, scatterplots, or other frequency plots, utilizing their proven visualization power. We have provided typical examples of multi-attribute visualization—such as 2-D transfer function specification and flow visualization—to demonstrate the applicability of our approach. The difference to discrete scatterplots is especially visible for low-resolution data sets and for data sets defined on grids with largely different cell sizes.

The main advantage of continuous scatterplots is that they are directly designed for input data defined on continuous domains. Therefore, this paper adds one missing piece to the general approach of ap-

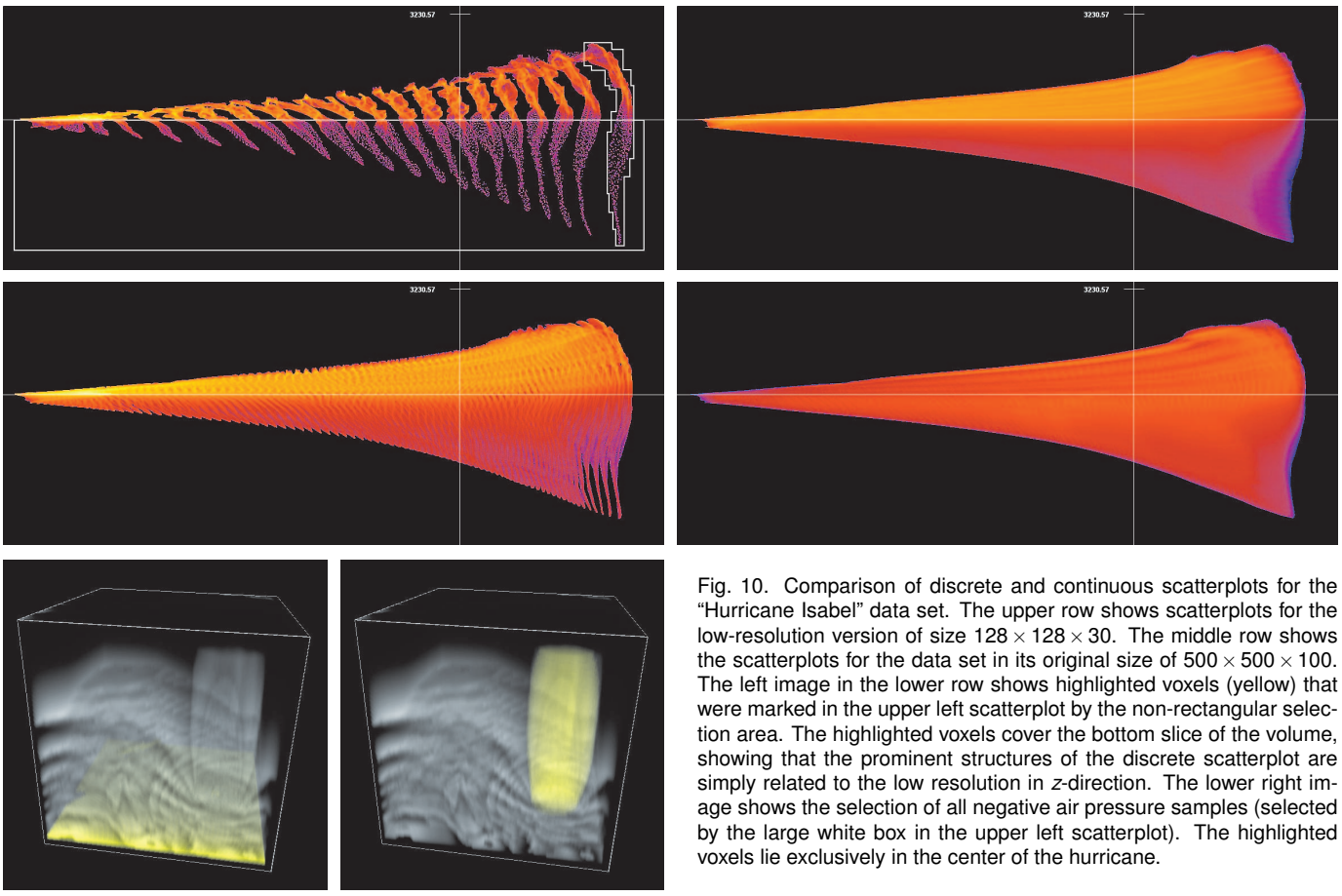

Fig. 10. Comparison of discrete and continuous scatterplots for the "Hurricane Isabel" data set. The upper row shows scatterplots for the low-resolution version of size $128 \times 128 \times 30$. The middle row shows the scatterplots for the data set in its original size of $500 \times 500 \times 100$. The left image in the lower row shows highlighted voxels (yellow) that were marked in the upper left scatterplot by the non-rectangular selection area. The highlighted voxels cover the bottom slice of the volume, showing that the prominent structures of the discrete scatterplot are simply related to the low resolution in z-direction. The lower right image shows the selection of all negative air pressure samples (selected by the large white box in the upper left scatterplot). The highlighted voxels lie exclusively in the center of the hurricane.

plying statistical and information visualization methods to scientific data. A related advantage is that continuous density plots scale well with increasing data set size, avoiding the issue of overplotting, which typically arises for large scientific data sets. Furthermore, continuous scatterplots are parameter-free; they do not require parameters for bucket size or other a-posteriori construction of density from discrete data. We also emphasize the advantages of the rigorous and generic mathematical model introduced in this paper. Not only does this model provide a solid and reliable basis for many variants of frequency plots of continuous data, but it also allows us to assess the errors introduced by previous discrete frequency plots, which can be viewed as examples of numerical approximation of continuous scatterplots. Finally, generalization has value of its own in any scientific discipline that strives for unification and simplification; in this case, generalization allows us to better understand and unify different frequency plot approaches.

One open question is whether and how 2-D scatterplots can be computed efficiently for non-linear interpolation. In particular, a direct computation for uniform or rectilinear grids (with trilinear interpolation) could be investigated in future research. We also plan to apply our approach to other application examples, including multi-modal medical-imaging data. Finally, the numerical errors introduced by discrete frequency plots (compared with continuous scatterplots) could be evaluated quantitatively; for example, numerical differences could be computed for large classes of data sets and resolutions to assess where continuous scatterplots would have the largest impact.

Acknowledgements

The "blunt-fin" data set is courtesy of the NASA Advanced Supercomputing (NAS) Division. We thank Roger Crawfis for providing the "tornado" data set. The "Hurricane Isabel" data was produced by the Weather Research and Forecast (WRF) model, courtesy of NCAR and the U.S. National Science Foundation (NSF).

References

[1] P. Billingsley. *Probability and Measure*. Wiley-Interscience, 3rd edition, 1995.

[2] H. Carr, B. Duffy, and B. Denby. On histograms and isosurface statistics. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1259–1266, 2006.

[3] J. M. Chambers, W. S. Cleveland, and P. A. Tukey. *Graphical Methods for Data Analysis*. Duxbury Press, 1983.

[4] R. Courant and D. Hilbert. *Methods of Mathematical Physics*. Wiley, 1962.

[5] H. Doleisch, M. Gasser, and H. Hauser. Interactive feature specification for focus+context visualization of complex simulation data. In *Proc. EG / IEEE Symposium on Visualization (VisSym)*, pages 239–248, 2003.

[6] H. Federer. *Geometric Measure Theory*. Springer, 1996.

[7] H. Goldstein. *Classical Mechanics*. Addison-Wesley, 2nd edition, 1980.

[8] D. H. Griffel. *Applied Functional Analysis*. Dover Publications, 2002.

[9] R. L. Harris. *Information Graphics: A Comprehensive Illustrated Reference*. Oxford University Press, 1996.

[10] Y. Ioannidis. The history of histograms (abridged). In *Proc. Very Large Databases (VLDB)*, pages 19–30, 2003.

[11] J. Kniss, G. Kindlmann, and C. Hansen. Multi-dimensional transfer functions for interactive volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 8(3):270–285, 2002.

[12] A. A. Saveliew, S. S. Mukharamova, N. A. Chizhikova, R. Budgey, and A. F. Zuur. Chapter 19: Spatially continuous data analysis and modelling. In A. F. Zuur, E. N. Ieno, and G. M. Smith, editors, *Analysing Ecological Data*, pages 341–371. Springer, 2007.

[13] C. E. Scheidegger, J. Schreiner, B. Duffy, H. Carr, and C. T. Silva. Revisiting histograms and isosurface statistics. *IEEE Transactions on Visualization and Computer Graphics*, 14(6), 2008.

[14] P. Shirley and A. Tuchman. A polygonal approximation to direct scalar volume rendering. *Computer Graphics*, 24(5):63–70, 1990.

[15] J. M. Utts. *Seeing Through Statistics*. Duxbury Press, 3rd edition, 2004.