DOI: 10.1111/cgf.14001
Eurographics Conference on Visualization (EuroVis) 2020
M. Gleicher, T. Landesberger von Antburg, and I. Viola
(Guest Editors)

Volume 39 (2020), Number 3

Sunspot Plots: Model-based Structure Enhancement for Dense Scatter Plots

T. Trautner, F. Bolte, S. Stoppel, and S. Bruckner

University of Bergen, Norway

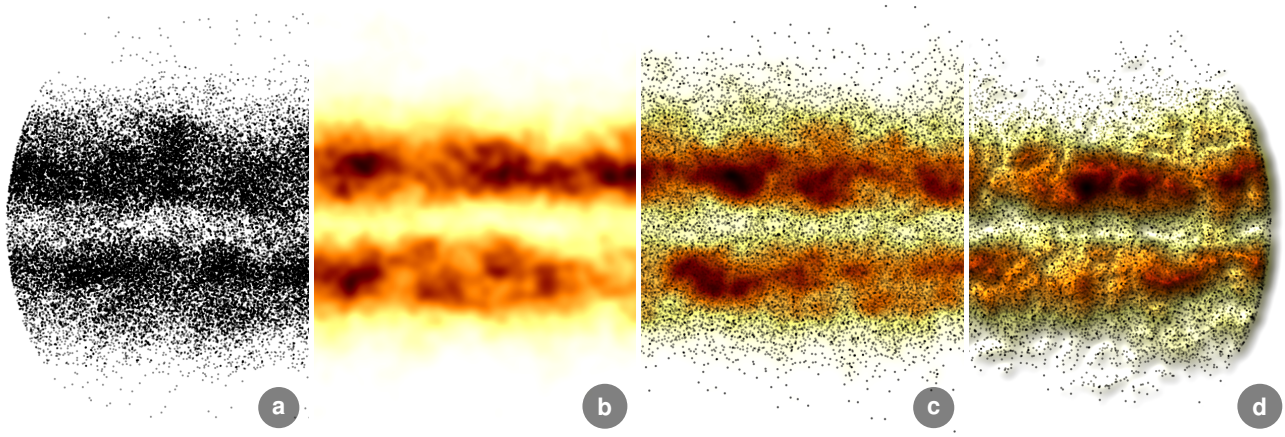

Figure 1: *Sunspot plots were inspired by the natural phenomenon of the same name visible on the sun's surface. These spots are colder and therefore darker, which is why they visually stand out from their bright surroundings. Between 1825-1867, the German astronomer Samuel Heinrich Schwabe documented the position of approximately 135,000 sunspots. When visualizing such dense data sets as scatter plots (a), it is often hard to estimate the density distribution of points due to overplotting. Density-based visualizations (b) are able to convey trends and clusters, but fail to effectively depict individual data samples. Our technique (c) is able to visualize density variations without abolishing the discrete nature of scatter plots. In addition, illumination can further emphasize local density changes, as shown in (d).*

Abstract

Scatter plots are a powerful and well-established technique for visualizing the relationships between two variables as a collection of discrete points. However, especially when dealing with large and dense data, scatter plots often exhibit problems such as overplotting, making the data interpretation arduous. Density plots are able to overcome these limitations in highly populated regions, but fail to provide accurate information of individual data points. This is particularly problematic in sparse regions where the density estimate may not provide a good representation of the underlying data. In this paper, we present sunspot plots, a visualization technique that communicates dense data as a continuous data distribution, while preserving the discrete nature of data samples in sparsely populated areas. We furthermore demonstrate the advantages of our approach on typical failure cases of scatter plots within synthetic and real-world data sets and validate its effectiveness in a user study.

CCS Concepts

• ***Human-centered computing*** → ***Visualization techniques; Information visualization; Empirical studies in visualization;***

1. Introduction

Scatter plots are a popular means of representation within various research fields. According to Friendly and Denis [FD05], it can be assumed that scatter plots represent 70-80% of the graphs used in scientific publications. Traditionally, scatter plots encode each data element with a marker placed on two ordered orthogonal dimensions. Simple yet versatile, scatter plots are able to effectively depict global and local regression patterns, clusters, and outliers in

the data. However, due to *overplotting*, scatter plots become less effective when dealing with an increasing number of samples. As a result, global and local data variations become increasingly hard to detect. Approaches like opacity regulation, reduction of marker size, or subsampling can increase the readability to some degree. Nevertheless, these approaches still fail when the number of data samples approaches the number of available pixels. Frequency-based visualization techniques such as density estimations offer solutions for such cases. These approaches interpret the data as samples of a continuous phenomenon that can be approximated by a model, e.g., a probability density function. A popular approach to approximate the probability density function is *kernel density estimation* (KDE). However, visualizations based on density functions are not well suited to depict local variations or singular points. Especially in low density regions, discrete encoding can depict singular data points better than density-based approaches.

In this paper we introduce *sunspot plots*, a novel visualization for bivariate scattered data. Sunspot plots are named after the astronomical phenomenon that they bear resemblance with, similar to the visualization of Schwabe's sunspot observations [ALG*13], shown in Figure 1. Our approach aims to alleviate the respective drawbacks of discrete and continuous representations. Instead of dismissing the discrete encoding of data points, we combine advantages of discrete and continuous data metaphors in one unifying visualization model. We interpret the probability density function as a surface over a two-dimensional field. Depending on the density of the underlying data, we create a blend between the continuous surface representation and the discrete point representation. Based on a GPU on-the-fly KDE, we are able to interactively display even large data sets with several thousands of data elements. Finally, using illumination sunspot plots can exploit the human mental model of surfaces to effectively convey the absolute density as well as local and global variations of the data. The contributions in this paper can be summarized as follows:

- A novel kernel density-based visualization technique that unifies discrete and continuous representation of large bivariate data
- A user study with several hundred participants in which we compared five different visualization techniques based on two representative tasks during the analysis of bivariate data

2. Related Work

Sarikaya and Gleicher [SG18] introduced four scatter plot design classifications depending on task and data characteristics: *point encoding*, *point grouping*, *point position*, and *graph amenities*. For reasons of clarity, the following section is divided into the first three categories, omitting graph amenities.

Point Encoding: Point encoding represents design decisions that affect the appearance of individual markers of data points. For example, color, size, opacity, blurriness, outline, or the selected symbol [SG18]. Li et al. [LvM10] provide studies concerning lightness, size, radii [LMv10], and contrast [LvM09] of different scatter plot symbols in relation to human visual perception. As part of our user study, we investigated, i.a., the effects of density-dependent color coding of individual points compared to conventional scatter plots and continuous heatmaps.

An example of point encoding are *bubble plots* [Pla05]. They display circles instead of points, whose diameter and color encode additional data-dependent properties. Unfortunately, scaling of circle radii simultaneously increases the chance of occluding neighboring circles. An ad-hoc approach to overcome overplotting is to uniformly downscale all radii in case of occlusions. Woodruff et al. [WLS98] extended this idea to a density-based approach that adjusts the size of each symbol depending on the density of its neighborhood. Mayorga and Gleicher [MG13] point out that when the data size exceeds the screen resolution, overplotting can no longer be avoided through scaling. This restriction intensifies further if the screen space is narrowed down, for example, by using *scatter plot matrices* [Har75] to simultaneously compare multiple scatter plots.

Apart from size, other visual properties such as color or brightness can be used to augment scatter plots. However, according to Trumbo [Tru81], color blending is limited by the nonlinear color perception of the human visual system which may vary even among people who are not considered as color blind. This may handicap the interpretation of quantitative data. An alternative would be transparency [Wil05, WB04, MAF15]. *Alpha blending*, for example, represents an implicit density encoding since points are rendered semi-transparently instead of completely opaque [The06]. Figures 2a and b show a comparison between a scatter plot with completely opaque points and one using alpha blending. Nevertheless, blending is limited to a few layers and it soon becomes difficult to distinguish between varying densities. This inspired us to include a mental model of the density distribution as surface representation in our visualization, in contrast to pure point blending.

Point Grouping: Point grouping is a design concept that aggregates multiple objects and therefore abstracts data into combined local regions. A typical example is *binning* [HDS*10], which reduces the number of elements to prevent overplotting. On the one hand, it supports the identification of correlations and data characteristics, but on the other hand, it hinders the location and selection of individual data points. Finding an adequate bin size can be seen as an analogue problem to finding suitable KDE kernel parameters.

HexBin scatter plots [CLNL87] are a representative example of a point grouping approach. They represent hexagon-shaped bins whose size depends on the enclosed number of observations. Dupont and Plummer [DP03] emphasize that hexagonal bins are more densely packed and reduce disruptive horizontal and vertical artifacts which could arise from square bins. A hybrid approach visualizing density estimations through binning as well as individual data points are *Varebi plots* [HMS97]. They prevent occlusions using overplotting indices per bin depending on the glyph shape, screen size, and data distribution. Another example are *sunflower plots* [CM84], a combination of hexagonal bins and sunflower-like glyphs where the total number of flower petals corresponds to the number of binned data points. To summarize even larger numbers of elements, the background color of each bin can indicate a multiplication factor for the petals [DP03]. The main disadvantage of binning approaches is the actual displacement and distortion of the underlying density locations [DWA10]. Another disadvantage of quantizing density through aggregated bins is the absence of perceptually smooth contours. Therefore, Mayorga and Gleicher [MG13] introduced *Splatterplots*. They combine smooth

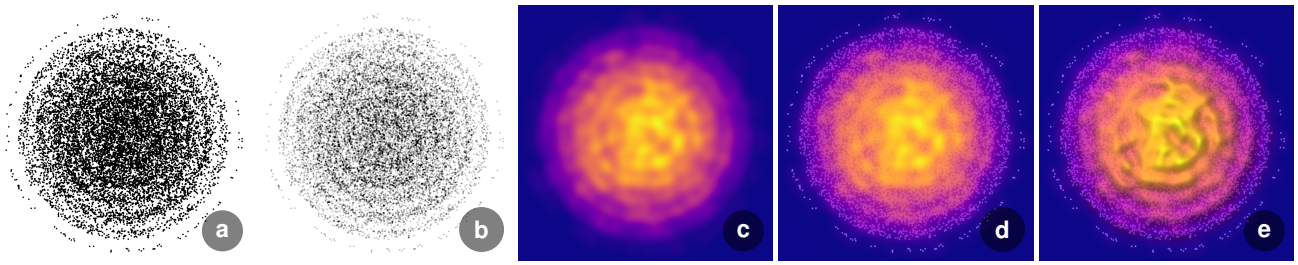

Figure 2: *Overview of currently well-established visualization techniques for bivariate data, which lay the foundation for our sunspot plots: (a) Classical scatter plots that can be improved using (b) alpha blending. (c) A visualization of a KDE using a continuous heat map representing the foundation of (d) sunspot plots depicting density information and singular data points (e) including the possibility to emphasize local variations using additional illumination cues.*

polygonal shapes enclosing dense regions, color blending for interrelations, and the explicit representation of subsampled outliers. Unfortunately, color blending could degenerate and synthesize new colors that are difficult to interpret. In contrast to color blending, we want to emphasize the perception of *shape from shading* [Ram88] as a possible alternative. In our study, we therefore examined how typical tasks performed with scatter plots are affected by shading. Similar to Splatterplots, we use an on-the-fly KDE. Point aggregation based on KDE has proven to be a solid and elegant approach which neither destroys the layout of the underlying data nor limits the number of data points. It is therefore used in many related applications, for example, density-based node aggregation used for large graphs [ZBDS12], visualization of graphs as continuous fields [vd03], visualization of high-density flows of streaming points [LBH18], or the emphasis on edges using line kernels [DLH11]. KDE is therefore not only a popular tool in statistics but also applicable in visualization.

Point Position: Design decisions regarding point position include rearranging or reprojecting the spatial position of individual points, reducing data through subsampling, or visualizing temporal changes as *animations* [CEJ*18, MM18]. Unfortunately, the time required to play an animation increases with the number of overlapping samples and it may be necessary to play the animation multiple times. Furthermore, Carr et al. [CLNL87] highlight that animation can help in detecting changes, but often visualizations of density differences are sufficient.

Another possible approach to prevent overplotting is through *subsampling*, in order to reduce the total number of underlying data points. A detailed analysis of clutter reduction techniques including subsampling is presented by Ellis and Dix [ED07]. Simple random sampling [ED02, DE02], for example, does not require any prior knowledge of the data set but can be disadvantageous since local or global trends may disappear, sampling artifacts could be introduced, or important outliers may not be preserved. More advanced approaches are *adaptive* or *stochastic subsampling* [CCM*14, BS06]. These approaches aim to reduce the density in highly occluded regions while preventing further decrease in density of sparse regions, which would result in information loss. Examples are *best uniform sampling* and *non-uniform sampling* [BS06] which are both based on perceptual user studies, and *multi-class sampling* [CCM*14, CGZ*19] that preserves point distributions. Ellis et al. [EBD05] present a user-controlled sampling lens enabling different sampling rates inside and outside the lens. We consider such *focus+context* functionality to be complementary and provide this functionality through an optional *magic lens* in our visualization. Figure 2c shows the result of a KDE visualized using a continuous heat map, and Figures 2d and 2e depict sunspot plots, without and with illumination cues, respectively.

Other possible approaches to reduce overplotting are *displacement techniques* [KH98, Kei00, RGE19]. Their advantage is that all individual points remain visible to the user. An example is *circular pixel placement* [KHD*09]. In case of an occlusion, the circular region around a data point's pixel position is searched and subsequently changed to the next free pixel closest to the original position. This approach has been further developed by Janetzko et al. [JHM*13] who introduced *ellipsoid point placement*. It performs an initial clustering and *principal component analysis* to generate ellipses whose orientation and aspect ratio visualize the local correlation of the data. Another possibility is adding *jitter* or marginal *random noise* [CCKT83, TGC03]. Instead of placing data points exactly above each other, minimal variations along the x and y axes are added. Disadvantages of these displacement techniques are the still limited number of available pixels on screen and the falsification of the original point positions which may obscure important data properties. Additionally, arbitrary visual patterns independent of the data, e.g., merging circular structures, are introduced.

Filtering operations [TGC94, BCLC97] aim to reduce the underlying data to the most relevant, meaningful, or significant selections through user input. Filtering already requires a certain basic knowledge of the data, thus is not suitable for untrained users. A representative example is the *magic lens filter* [SFB94]. Similarly, *distortion techniques*, such as zooming operations (uniform) or *fish-eye views* [BGR06] (non-uniform), allow decluttering and a focus on interesting data regions. Tominski et al. [TGK*17] provide a detailed survey of visualization techniques using lenses. Carpendale et al. [CCF95] emphasize the importance of informing the user about the spatial distortion, for example, by using superimposed grid lines or adequate shading techniques. Similar to filtering, these techniques are user-dependent and therefore particularly affect inexperienced or untrained users. This is why we prefer either automatic scatter plot designs [MPOW17] or visualizations with generally chosen parameters applicable to various use cases.

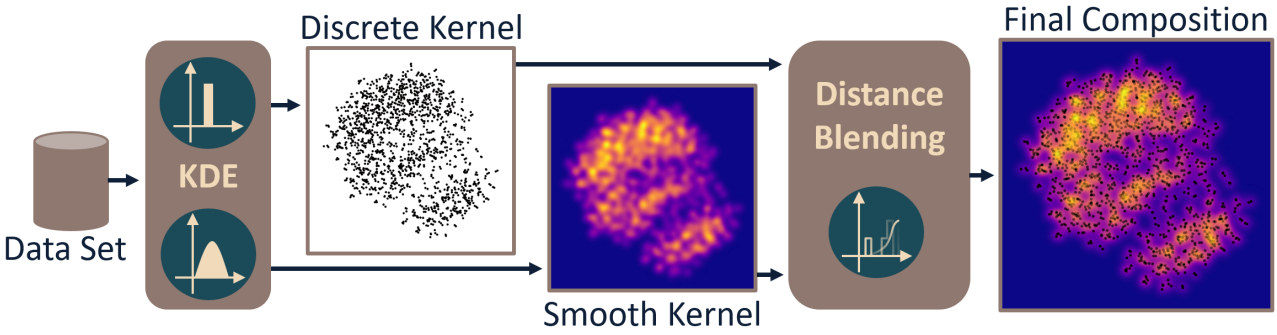

Figure 3: *Overview of our visualization system pipeline: Initially, we compute a discrete and a continuous representation of the data. Subsequently, both representations are composed into a unified visualization emphasizing continuous and discrete phenomena.*

Finally, we want to discuss multidimensional approaches. Dang et al. [DWA10] present *stacked plots*, a three-dimensional visualization approach that stacks data points in case of occlusions. Unfortunately, adding a third dimension introduces further problems such as occlusions depending on the viewing direction, perspective distortions, and difficulties in interpreting size and distance which may require additional user interaction. A multidimensional focus+context approach is presented by Staib et al. [SGG16]. They introduce implicit filtering through continuous *depth-of-field* techniques. Bachthaler and Weiskopf [BW08, BW09] present *continuous scatter plots*, a parameter-free continuous density plot that is capable of visualizing multidimensional input data defined on continuous domains. Similar to our approach, continuous density plots scale well with increasing data size, preventing problems with overplotting, but do not preserve the discrete data properties in sparse regions. Lastly, we would like to draw attention to the inspirational work of Sprenger et al. [SBG00], which demonstrated the usefulness of hierarchical clustering with the help of implicit surfaces, and the work of van Wijk and Telea [vT01], which enriched visualizations of scalar functions of two variables by adding ridges.

3. Sunspot Plots

Overplotting is a typical problem of scatter plots even when displaying only moderate amounts of data, as shown in Figures 1a and 2a. A common strategy to increase the readability of scatter plots is opacity reduction of individual points. We show results of this strategy in Figure 2b – although some parts of the data are depicted clearer, overplotting is still present. Frequency-based visualization techniques are a common solution for large scattered data sets. Typically, density distributions are conveyed via *heatmaps* [SG18]. As shown in Figure 2c, these are not very well suited to depict small variations or outliers. Since the applied color has a strong influence on the perceived change of the density function, it is important to choose a perceptually uniform color map such as *Plasma* [Hun07].

The goal of our approach is to support the exploration of bivariate scattered data by providing the best suited visual encoding depending on the local data density. We furthermore aimed for a technique that can be easily integrated into already existing visualization systems. One key aspect of our technique is an importance function α that regulates a smooth transition between a discrete encoding in sparse regions and a density-based surface encoding in data-rich regions. Figure 3 shows a general overview of our method. We simultaneously interpret the data set as a discrete and a continuous phenomenon. The discrete aspects are depicted via a conventional scatter plot. In order to capture the continuous aspects of the data, we interpret the underlying data distribution as a two-dimensional field that we approximate through a model. While there are countless models to approximate the data distribution, we use an on-the-fly KDE to achieve interactive performance. In the final step of our pipeline, the discrete and continuous models are blended based on the importance function α. In the following, we provide a detailed description of our approach.

3.1. Kernel Density Estimation

Bivariate data points can often be seen as samples of a continuous phenomenon, i.e., a *probability density function* (PDF). Assuming the data is described through a PDF, the data distribution can be modeled through *kernel density estimation* (KDE). KDE is a non-parametric approach to approximate the PDF from a finite number of samples, where the PDF model $\hat{f}(x)$ at point x is computed through addition of kernels at the positions of each data sample:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \tag{1}$$

$$K_{disc}(u) = \begin{cases} 1, & \text{if } |u| \leq 1 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$K_{cont,\sigma}(u) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{u^2}{2\sigma^2}}, & \text{if } |u| \leq 1 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

wherein K is the kernel function, h is the kernel bandwidth, n is the number of the individual samples x_i, and σ defines the width of the Gaussian kernel. *Silverman's rule of thumb* [Sil86] suggests a globally optimized bandwidth parameter h, but due to the strong data and task dependency, we nevertheless allow h as well as σ to be adjusted by the user. In addition, it is possible to have different bandwidths h for K_{disc} and K_{cont}. Naturally, the properties of the reconstructed PDF highly depend on the kernel chosen. Conventional scatter plots, for example, can be seen as reconstructed models using discrete kernels with constant intensity, as sketched

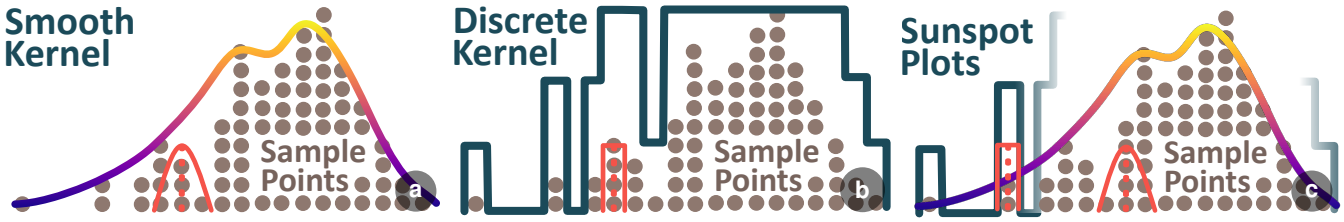

Figure 4: *(a) Smooth surface representations are well suited to depict the general data distribution, but badly suited to depict single data points. (b) Discrete encodings are opportune for the visualization of singular data samples but can reach an opacity threshold and exhibit overplotting. (c) We utilize the strength of both approaches to convey the data probability function in dense and sparse regions.*

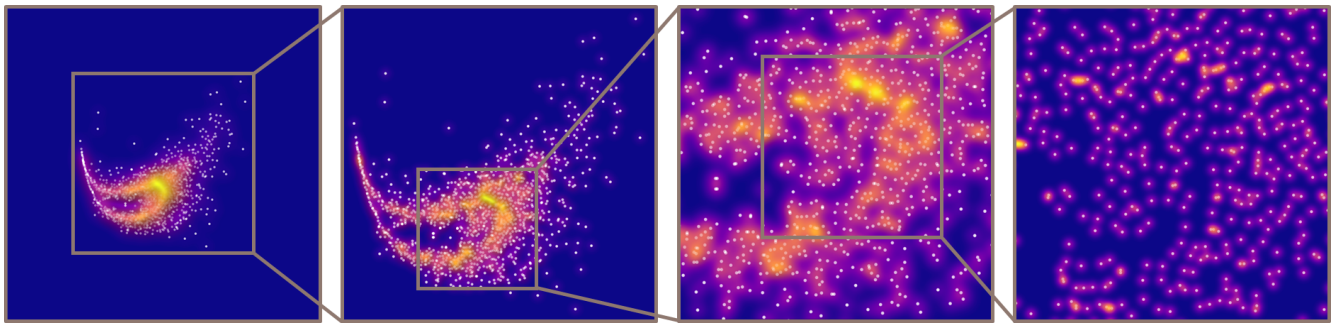

Figure 5: *The aggregation level of sunspot plots automatically adapts to the zoom level. From left to right we zoom into the highlighted areas. It can be seen that the aggregation becomes more and more detailed as the zoom level increases.*

in Equation 2. Smooth kernels, on the other hand, create continuous surfaces that are well suited to describe patterns in the data distribution. The normal distribution function, as described in Equation 3, is a common choice for a continuous kernel. As we illustrate in Figure 4, not all kernels are equally suited for all data distributions. Smooth kernels, such as normal distribution kernels, are better suited to reconstruct the data distribution within dense data regions [MG13]. On the other hand, discrete kernels are well suited for low density areas, as they are better able to convey outliers and their position than continuous representations. In our approach, we aim to reduce the visual complexity of the visualization by smoothly changing the kernel choice based on the local data complexity instead of adapting the bandwidth. In contrast to Splatterplots [MG13], we do not make a strict separation between the discrete and aggregated data interpretation. Instead, sunspot plots smoothly blend between the continuous and discrete representation based on an importance function α. There are various possible candidates for such an importance function. To reduce the computational complexity, we construct α by normalizing the already computed PDF $\hat{f}$ at point x as:

$$\alpha(x) = \frac{\hat{f}(x)}{max\left\{\hat{f}(t) \mid t \in \mathbb{R}\right\} - min\left\{\hat{f}(t) \mid t \in \mathbb{R}\right\}}. \quad (4)$$

As a result, α can be directly used as opacity while blending. This way, dense regions are aggregated to a continuous representation, while regions with low density are depicted through a discrete representation. The typical exploration of dense scatter plots requires *panning* and *zooming* in image space. When zooming, the ratio between the data samples and available image space changes. Our approach inherently accounts for this by fixing the kernel bandwidth within the image space. Hence, we smoothly transition the visual encoding from a surface representation to a discrete plot. An example of how the aggregation becomes more detailed as we zoom in on the data is shown in Figure 5.

3.2. Surface Properties

As mentioned earlier, the probability density function over a two-dimensional field can be naturally interpreted as a surface. The field of *computer graphics* has a long tradition in effectively rendering three-dimensional surfaces. It is therefore reasonable to investigate well-established techniques for truthful and easy-to-understand surface representations. Already in 1950, Gibson [Gib50] discussed aspects that affect our shape perception of surfaces. These attributes have a direct relationship to data properties captured with a PDF:

Distance to surface relates directly to the surface height, i.e., absolute values of the probability density function. When using surface representations for bivariate data, a color encoding is the most effective way to convey absolute function values [TSW*07].

Quality and orientation of slant, or slope, corresponds to changes of the PDF. Gibson [Gib50] furthermore mentions the shape at a given slant, i.e., the curvature of the surface, as an important visual cue. The curvature depicts the second derivative of the PDF. Both, the slope and curvature, are conveyed effectively using the *Phong illumination model* [Pho75].

Impression of contours arises when surface parts are separated from each other or when surface areas are not well accessible to

light. The nature of the PDF does not allow for discontinuity of surfaces, therefore we only need to account for surface accessibility. A measure that is equivalent to surface accessibility is the measure of *ambient obscurance*, which is the amount of ambient light that is stopped from reaching a point. Ambient obscurance, first introduced by Zhukov et al. [ZIK98], is a technique for natural-looking lighting effects with a small fraction of computational costs, compared to global illumination. In other words – obscurance measures, for each point on the surface, the degree to which its hemisphere is obscured by neighboring surfaces. In this paper we use scalable ambient obscurance, introduced by McGuire et al. [MML12], to efficiently approximate the ambient obscurance of the surface.

Naturally, illumination affects the colors of the rendered surface, with slightly darker colors in shadow regions and brighter colors in areas facing light. However, since the human perceptual system is processing global structures and not only the local color information, as for example shown by the Adelson's *Checker-Shadow Illusion* [TM19], we postulate that using shading with sunspot plots will not have a negative effect on the perception of the surface properties, but might even yield an advantage over solely color-based sunspot plots. This assumption and other hypotheses were tested in an extensive user study which is described in detail in Section 6.

3.3. Blending Function

As already mentioned, color is the most effective way to convey the surface height or, in our case, the absolute PDF values. To emphasise the discrete nature in sparser regions, the discrete kernel is using a constant color. A natural question that arises when combining two visual stimuli is the choice of the blending function. Following the reasoning of the importance function α, as sketched in Equation 4, we blend the two stimuli using the *Porter-Duff* [PD84] *over* operator. The blending of **A over B** is defined as follows:

$$C = \frac{1}{\alpha_C}\left(\alpha_A A + (1-\alpha_A)\,\alpha_B B\right) \tag{5}$$

$$\alpha_C = \alpha_A + (1-\alpha_A)\,\alpha_B \tag{6}$$

wherein C is the final pixel color, A the continuous, and B the discrete stimuli. In our case, α_A corresponds to the importance function α as computed from the normalized PDF and α_B can be considered as constant value 1. The background color can be chosen arbitrarily, but should preferably be either a neutral color, such as white, or the color depicting the lowest value in the density plot. We chose the lowest density value for all of the Figures shown in this paper.

4. Implementation

Our aim was to implement sunspot plots as interactively as possible, allowing the user to adjust individual parameters, such as the bandwidth h or the width of the Gaussian kernel σ, in real-time. The underlying data should thus be easy to explore and interpret. In order to achieve this, the presented approach was implemented as a framework based on C++ and OpenGL. Apart from the most necessary operations, such as loading and processing *comma-separated values* (CSV) files as input, all calculations were written in GLSL and performed in parallel on the GPU. An imported CSV file represents a table of data values from which the user can select two columns. This selection is done using *ImGui* [Cor19], a *graphical user interface* (GUI). These columns are then mapped to the x and y axis. In addition, the GUI provides sliders to dynamically adjust parameters such as opacity of the surface model, scale of the samples, or the used bandwidth of the KDE. The visualization is then immediately updated and provides visual feedback based on the changed parameters. This allows the user to analyze diverse data sets and generate expressive visualizations. In the following, we will present our implementation consisting of four *render passes*.

First render pass: We start our visualization pipeline by rendering a scatter plot using the discrete kernel K_{disc}. Therefore, all scattered points are positioned according to their x and y coordinates, scaled and colored depending on user settings. While rendering each point, we additionally perform an implicit opacity modulation using an *additive blend function*. To prevent rasterization artifacts while splatting, we further perform *anti-aliasing* by smoothly fading out the outlines of the sample points.

(Optional) Pilot KDE: This optional computation is performed only if the user prefers an adaptive KDE with local bandwidth instead of a KDE with constant global bandwidth. In case of an adaptive KDE, a comparably smaller kernel is used in regions with higher density than in sparser regions. This adaptability requires a pilot KDE with constant bandwidth which can be used to adapt the kernel depending on the estimated density from the pilot study.

Second render pass: This render pass performs the continuous (adaptive) KDE, which is similar to the computation from the first render pass. They differ, however, in that now the radii of the points are increased. This is necessary for a smooth evaluation of K_{cont}. As usual, the increased radii represent the bounding geometry of the Gaussian kernels for a fixed cutoff value where their contribution is considered to be zero. Next, the density function is evaluated per pixel. Again we use additive blending, but this time to sum up the density contributions of individual fragments.

Third render pass: Based on the results of the second render pass, we now calculate the actual surface model. Here we take advantage of the fact that the height of the surface directly corresponds to the accumulated density values. During this rendering stage, we additionally store the minimum and maximum values of the calculated PDF which is necessary for its normalization.

Fourth render pass: Here we apply the perceptually uniform Plasma color map from *Matplotlib* [Hun07] to the calculated density range and blend it with the scatter plot using the above described blending and importance function. Optionally, local Phong illumination [Pho75] and ambient occlusion [MML12] can be additionally applied. When using illumination, a single point light is positioned "top-left" according to the recommendations from Wambecke et al. [WVBT16] using an azimuth of 120°, elevation of 45°, and the five-fold distance of the diameter of the object in the center.

We plan to make our stand-alone framework available to the community under a permissive license to allow for extensions and to simplify the integration into existing applications.

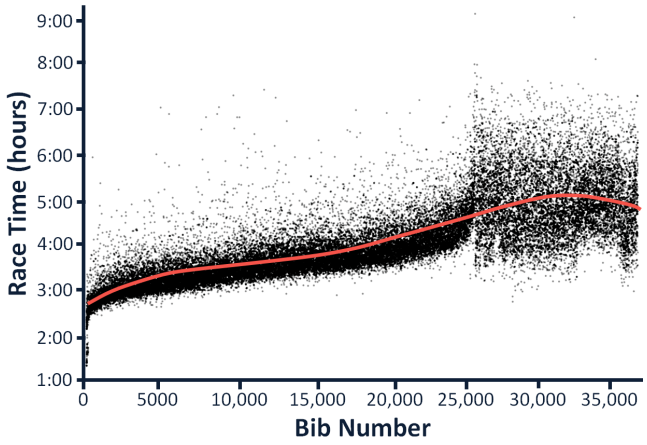

Figure 6: *Scatter plot of the 2014 Boston Marathon including a red trend line of degree four, calculated using the method of least squares. Notice the clearly different point distribution and trend of the last starter wave, beginning with the bib number 27,000.*

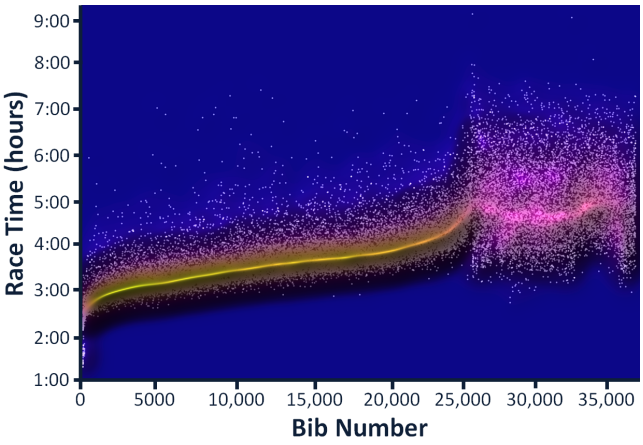

Figure 7: *Example of a sunspot plot that indirectly approximates the best fit polynomial visible as ridge line between bib number 0 and 25,000, and a race time between 2 and 5 hours.*

5. Usage Examples

In this section, we demonstrate the strength of sunspot plots based on three real-world data sets from different scientific fields, illustrating the applicability of our approach for large data sets with varying data density. All examples show possibilities of how global trends or clusters and individual points or outliers can be explored. For an easier comparison, we again used the *Plasma* color map [Hun07] in all our examples.

5.1. Boston Marathon

As our first example, we will analyze the 2014 Boston Marathon data set [Mil20] including the results of approximately 32,000 participants. A classical scatter plot using alpha blending including a polynomial trend line is shown in Figure 6. The x-axis of the plot encodes the unique *bib number*, which is assigned to each participant depending on the qualification time. The faster he or she was, the lower the number and the earlier the marathon starting time. The y-axis encodes the total time in hours that the participant needed to finish the complete marathon.

Next, we will visualize the same data set using sunspot plots. The result is shown in Figure 7. Similar to scatter plots but in contrast to classical heatmaps alone, individual sample points from sparse regions are preserved. An essential example is the group of runners with the lowest bib numbers and a race time between 1 and 2 hours. These are the participants of the *wheelchair race* including the South African participant Ernst F. Van Dyk who won the race with a time of 1:20:36. In this example, sunspot plots indirectly show the approximation of the non-linear trend using their density encoding. Comparing Figure 6 and Figure 7, it becomes visible that sunspot plots clearly show that the trend can be well approximated between a bib number of 0 and 27,000 but is only fuzzy and less meaningful between 27,000 and 36,000.

5.2. World Cities

As second example, we will analyze a data set obtained from the GeoNames database [Bou19]. This data set contains the geo-spatial locations of cities with more than 500 inhabitants worldwide accounting for 189,280 entries in total. The geo-spatial information can be used to explore the agglomerations of cities and identify areas that may require infrastructure extension or reveal environmental conditions that influence the cities' development, such as mountain chains or valleys. In the following, we will focus on insights that can be revealed using sunspot plots. Figure 8 shows a close-up of central Europe including several highlighting structures.

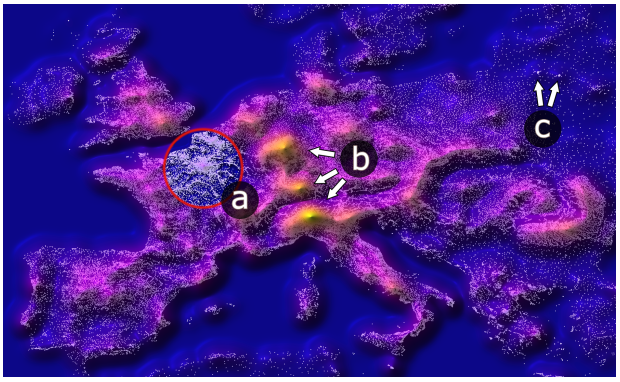

Figure 8: *Geo-spatial data of cities with more than 500 inhabitants in Europe. We have annotated a typical use case of a magic lens (a), densely populated metropolitan areas (b), and examples of sparse features (c) which are all discussed in detail in Section 5.2.*

Next, we will discuss the highlighted examples from Figure 8. Annotated by (a), a practical use case of the magic lens becomes apparent, revealing the exact catchment areas and suburbs of Paris. It uncovers inhabited areas and the resulting large-scale road networks in northern France. At letter (b), we can observe larger clusters around Frankfurt, Zurich, and Milan. It simultaneously shows

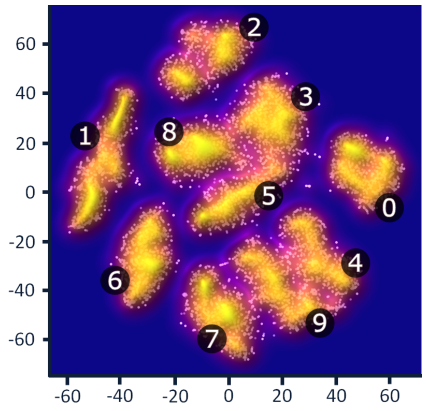

Figure 9: *Visualization of the t-SNE result of handwritten digits using sunspot plots. We have additionally annotated the implicitly created clusters with the corresponding numbers from 0 to 9.*

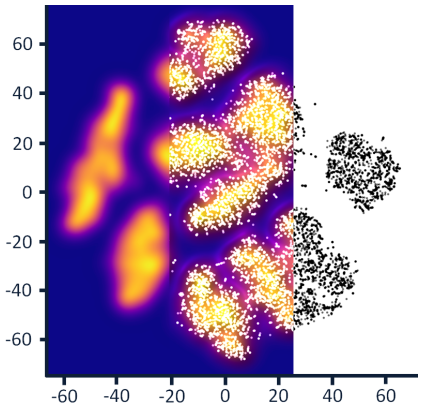

Figure 10: *Overview of three different visualization techniques: (left) a heat map that neglects individual samples, (middle) naive blending of heat map and scatter plot where overplotting hides the actual density encoding, and (right) a scatter plot alone which complicates the detection of cluster centers.*

that the cluster around Frankfurt is more widely populated than the concentrated area around Zurich. Because of the Alps, the area between Switzerland and Italy is only sparsely populated. The ridge-like cluster in Italy is particularly elongated and extends from Turin in the west to Venice in the east. The two arrows from letter (c) point to the largest cities in Lithuania, Vilnius and Kaunas. Both are, in contrast to pure density encoding, clearly visible although the surrounding country is evenly but very sparsely populated.

5.3. t-SNE of Handwritten Digits

As third example, we will analyze the test set of the MINST database of handwritten digits [LCCJ20]. It contains 10,000 28-by-28 grayscale images of labels from 0 to 9. After an initial dimensional reduction from originally 784 to 50 dimensions, using a *principal component analysis* (PCA), their dimensions are further reduced to two using the *t-SNE Barnes-Hut algorithm*. The resulting sunspot plot of similar point clusters is shown in Figure 9.

In the following, we want to justify our design decisions by investigating how sunspot plots from Figure 9 would change if one were to naively blend the computed heat map with a scatter plot. This comparison including a heat map and scatter plot alone is shown in Figure 10. It seems that there are already occlusions that complicate the interpretation although this data set is comparatively small. It is no longer possible to perceive the density of regions that contain sample points, but only that of regions without. It is therefore no longer possible to interpret the actual underlying data. This negative effect would be even worse if more points were to be visualized.

6. User Study

One of the main concerns with sunspot plots is that the blending of individual points and density representations using an importance function might still negatively influence the user perception of absolute and relative density values. We therefore conducted a controlled user study to analyze the human visual perception dealing with dense scatter plots, overplotting, and density estimations. We compared well-established techniques like scatter plots using common opacity modulation or color encoding, continuous heatmaps, and our approach in a regular and an illuminated alternative. In total, we analyzed five different visualization stimuli, as shown in Figure 11. The used questionnaires and data collected during the study are included as supplemental material.

6.1. Hypotheses

We utilize the classical scatter plot with alpha blending as base line for density encoding visualizations. One could argue that the easiest way to encode density would be to color individual sample points by their specific density value. Sarikaya and Gleicher [SG18] suggest that *point encodings*, as in colored scatter plots, would not improve density comparison tasks. Another approach would be *point grouping* as in continuous heatmaps. According to Sarikaya and Gleicher [SG18], this would improve the perception of density values and hence support comparisons. We therefore use colored scatter plots to gather empirical evidence for the statement that *point groupings* provide better support than *point encodings*. *Point groupings*, however, hide individual sample points. Since sunspot plots combine scatter plots and heatmaps, we aim to examine whether the blending of discrete points and a continuous heatmap introduces drawbacks in density estimation tasks. Finally, we aim to investigate if the introduction of illumination improves the perception of surface properties or if it leads to undesired side effects. We therefore state the following hypotheses:

H1: Density-dependent color coding of individual scatter plot points increases the users' accuracy of estimated density values compared to scatter plots with alpha blending alone.

H2: Sunspot plots do not perform worse than continuous heatmaps for absolute and relative density estimations.

H3: Applying an illumination model in combination with ambient occlusion shading to sunspot plots improves the users' accuracy when estimating density values.

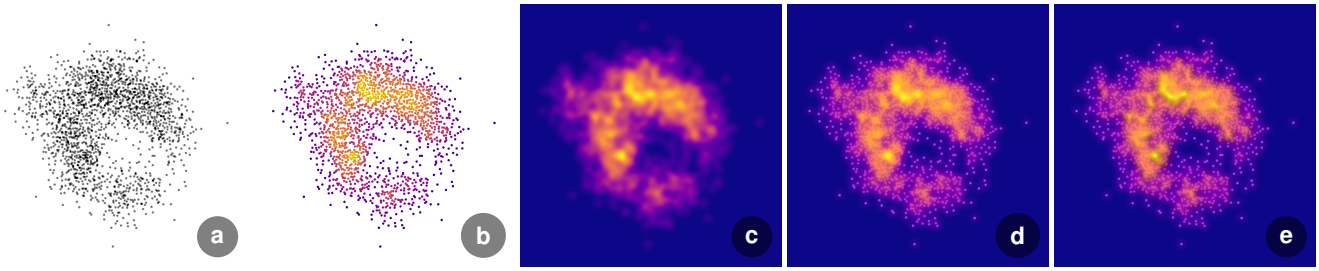

Figure 11: *Overview of the five visualization stimuli used in our user study: (a) A classical scatter plot with opacity modulation to encode overplotting. (b) A scatter plot with density-dependent color coding of individual points. (c) A smooth density representation using a continuous heatmap. (d) Sunspot plots and (e) sunspot plots including Phong illumination [Pho75] and ambient occlusion shading [MML12].*

Stimuli:	SP vs. **CSP**	CSP vs. **CHM**	CHM vs. SSP	SSP vs. SSSP	CHM vs. **SSSP**
Participants:	108 (56, 52)	106 (52, 54)	108 (54, 54)	98 (54, 44)	98 (54, 44)
Assign	$p = 2.81 \times 10^{-4}$ *	$p = 1.15 \times 10^{-13}$ *	$p = 0.161$	$p = 0.251$	$p = 0.014$ *
Posthoc Test	$\widetilde{m}_{SP} = -16$	$\widetilde{m}_{CSP} = -6.5$	$\widetilde{m}_{CHM} = 1$	$\widetilde{m}_{SSP} = 1$	$\widetilde{m}_{CHM} = 1$
(*Wilcoxon*):	$\widetilde{m}_{CSP} = -6.5$	$\widetilde{m}_{CHM} = 1$	$\widetilde{m}_{SSP} = 1$	$\widetilde{m}_{SSSP} = 0$	$\widetilde{m}_{SSSP} = 0$
Compare	$p = 1$	$p = 1.54 \times 10^{-7}$ *	$p = 0.663$	$p = 0.104$	$p = 0.271$
Posthoc Test	$\widetilde{m}_{SP} = 50\%$	$\widetilde{m}_{CSP} = 50\%$	$\widetilde{m}_{CHM} = 83.3\%$	$\widetilde{m}_{SSP} = 83.3\%$	$\widetilde{m}_{CHM} = 83.3\%$
(*Wilcoxon*):	$\widetilde{m}_{CSP} = 50\%$	$\widetilde{m}_{CHM} = 83.3\%$	$\widetilde{m}_{SSP} = 83.3\%$	$\widetilde{m}_{SSSP} = 66.6\%$	$\widetilde{m}_{SSSP} = 66.6\%$

Table 1: *Pairwise comparison of the study results of the assign task as well as the compare task using a Wilcoxon rank sum test. Stimuli names with superior performance in pairwise comparisons are shown in bold and significant p-values are marked with an asterisk (*).*

6.2. Experiment Design and Tasks

We designed an experiment in which participants are shown images of a given visualization technique and asked to answer one question regarding the data density per image. The study tested one independent variable, the visualization type, with five levels: *scatter plot* (SP), *colored scatter plot* (CSP), *continuous heat map* (CHM), *sunspot plot* (SSP), and *shaded sunspot plot* (SSSP). We followed a between-subject design to avoid a learning bias by assigning one visualization type to each participant.

The study consisted of two analysis tasks: *assign* and *compare*. We adhered to Sarikaya and Gleicher's list of *abstracted analysis tasks* [SG18] that are performed on scatter plots and considered one task from each of their defined task types as most relevant:

- **object-centric:** task 4 - *object comparison*
- **browsing:** task 5 - *explore neighborhood*
- **aggregate-level:** task 11 - *density comparison*

In the *assign* task, users had to estimate the density value from the bounded interval [0,100] in a highlighted region within the visualization. The correct value was computed using regular KDE, where a value of 100 corresponds to the densest point in the data and 0 to the sparsest. The *error* was computed as the difference between the participant's answer and the averaged per-pixel density within the highlighted region. In the *comparison* task, we highlighted three regions (A,B,C) and asked users which region contains either the highest or lowest density. Possible answers were A, B, C, and all three regions had the same density. We assured the participants that there was exactly one correct solution. The answers were therefore evaluated as a binary, correct or wrong.

Each participant was asked 12 questions, 6 per task type, each showing an image of a defined visualization type applied to a different data set. In order to minimize data bias, we selected a total of 84 data sets with varying densities from recent publications by Abbas et al. [AASB19] and Aupetit et al. [ASAB19], analyzing cluster patterns in scatter plots. In each data set we highlighted 4 different sets of density regions, 2 per task type, leading to a total of 336 images per visualization type. Each participant saw 12 of these images (2 each of sparse, moderately sparse, and dense data set in each of the two tasks). This ensures that on average, all data sets are displayed with similar frequency throughout the study. We randomized both the order of tasks and the order of questions within each task block to reduce potential learning effects. Participants provided answers via numeric text entry, had no time limit, and were told to come up with their best estimate. To assess the user performance we computed the *average error* for the *assign* task and the *average percentage of correct answers* for the *compare* task.

6.3. Participants and Procedure

The experiment was performed as an online user study using *Amazon Mechanical Turk*. We assigned 70 participants per visualization type resulting in a total of 350 participants, of which 260 (SP: 56, CSP: 52, CHM: 54, SSP: 54, and SSSP: 44) passed the attention checks. Participants (169 male, 90 female, and 1 with no answer) ranged from 18 to 68 years. We determined whether participants were wearing glasses (167 without glasses, 92 with glasses, and 1 with no answer) or were suffering from color vision deficiencies (245 without deficiencies, 8 reporting a deficiency, 6 reported that they do not know, and 1 with no answer).

At the beginning of the study, users saw a set of instructions based on their assigned visualization type. Next, they had to carry out two different task blocks (*assign*, *compare*), each consisting of 1 example, 1 practice trial, and 6 questions. When answering the practice trial, users were shown the correct answer including an explanation for how the answer was obtained. Each task block contained a simple attention check to filter out inattentive participants. At the end, participants had to answer general questions about the study and about themselves.

6.4. Study Results

We conducted a one-sample *Kolmogorov-Smirnov* test for each task type revealing that the underlying data are not normally distributed at the 5% significance level. We therefore used the *Kruskal-Wallis* non-parametric test indicating that the visualization stimuli differ significantly (*assign*: $\chi^2(4) = 171.35$, $p = 5.37 \times 10^{-36}$ and *compare*: $\chi^2(4) = 68.64$, $p = 4.38 \times 10^{-14}$). We rejected the null hypothesis assuming that all visualizations had similar effects and conducted *Wilcoxon* rank sum tests for pairwise comparisons. The detailed study results of the *assign* task and the *compare* task are shown in Table 1. Individual significant results from both task types were additionally highlighted.

Figure 12 displays the user *error* in density value estimations for the task type *assign*. The results show that users of CSP made significantly less errors than users of SP. This observation aligns with our hypothesis **H1** that a color coding of scatter plots can already improve the users' accuracy in density estimation tasks. Interestingly, we did not detect a significant difference in the *compare* task. While users of CSP performed better than users of SP, they still performed significantly worse than users of CHM in both the *assign* and *compare* tasks. This finding confirms previous results, for example from Sarikaya and Gleicher [SG18].

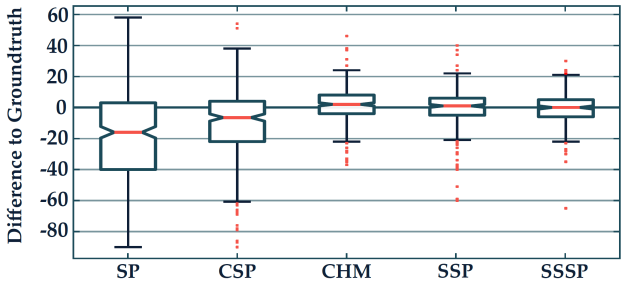

Figure 12: *Overview of the errors during the density value estimations for the assign task in all five visualization stimuli.*

As shown in Table 1, we could not find a significant difference between the user performance in CHM and SSP stimuli, neither for the *assign* task nor for the *compare* task. This aligns with our hypothesis **H2** that sunspot plots do not perform worse than continuous heatmaps for density estimation tasks. This provides us furthermore with an indication that our visualization method can indeed utilize the individual benefits of scatter plots and continuous heatmaps without losing the advantage in density estimations that heatmaps provide.

We could not confirm our hypothesis **H3**, that the introduction of an illumination model and ambient occlusion would improve the users' accuracy in estimation density values in sunspot plots. Users of the SSP and SSSP stimuli did not perform significantly differently, neither for the *assign* task nor for the *compare* task. However, as shown in Table 1, we saw a significant difference between users of CHM and SSSP, showing that shaded sunspot plots can provide a higher accuracy than continuous heatmaps for the *assign* task.

In summary, we can observe that blending continuous and discrete representations did not have a negative effect on the typical tasks for density-based visualizations, while having the advantage of including additional information of the individual data samples. A comparison of pure heatmaps with the illuminated sunspot plots in the *assign* task showed a significant performance improvement, however this result should be taken with a grain of salt, since no clear improvement could be observed between illuminated and non-illuminated sunspot plots. Hence, for the basic tasks we considered in this study, we can conclude that additional illumination likely does not provide substantial benefits, but we could not find evidence that it was harmful, either. An interesting topic for further investigation is whether this holds true for other types of tasks, e.g., ones that relate to patterns or high-frequency variations in density.

7. Performance

The following performance measurements were conducted on a desktop computer equipped with an Intel Core i7-8700K CPU (3.7 Ghz), 16 GB RAM, an NVIDIA GeForce RTX 2080 graphics card with 8 GB of texture memory, and Windows 10 Home 64-bit as operating system. In total, we analyzed nine data sets. Six of them were created artificially with different sampling rates (from 1,000 to 1,000,000 points). The other three data sets correspond to the usage examples from Section 5. During these tests, we scaled the data sets so that their bounding box corresponded to the exact viewport size. Data set *World Cities* is the only exception. Here we zoomed in on Europe to generate a more realistic application scenario. In total, we analyzed two different viewport sizes (768x768 and 1280x1280). Table 2 shows a detailed overview of the test scenarios including the number of sample points and the average number of frames per second (avg_{min}^{max}) per viewport.

Data Set	Points	Viewport 768x768	Viewport 1280x1280
Artificial Data	1,000,000	$11.51_{9.93}^{11.75}$	$4.33_{4.31}^{4.41}$
	200,000	$37.18_{36.90}^{37.32}$	$13.72_{13.31}^{13.95}$
	100,000	$55.70_{55.59}^{55.78}$	$21.45_{21.43}^{21.47}$
	50,000	$83.875_{83.48}^{84.22}$	$31.28_{29.28}^{31.77}$
	10,000	$116.35_{115.47}^{117.25}$	$38.06_{37.85}^{38.12}$
	1,000	$261.44_{257.06}^{266.76}$	$75.30_{74.69}^{75.87}$
World Cities	189,280	$34.01_{33.96}^{34.06}$	$13.60_{12.07}^{13.95}$
Boston Marathon	31,984	$68.31_{65.52}^{69.06}$	$26.14_{23.13}^{26.75}$
t-SNE	10,000	$128.01_{127.58}^{129.35}$	$49.12_{48.92}^{50.03}$

Table 2: *Overview of the analyzed test scenarios including their names, number of sample points, and the average framerate in frames per second (avg_{min}^{max}) using two different viewport sizes.*

8. Discussion and Limitations

We have developed sunspot plots as a flexible visualization approach for large scattered bivariate data. Although relatively straight-forward, we are unaware of previous work that smoothly combines discrete and continuous representations of scattered data in this manner. Furthermore, we implemented automatic adjustment of the kernel size to the current zoom level; we also support a non-uniform kernel computation based on the density distribution, and allow for the use of a magic lens that shows more local details by reducing the kernel bandwidth. Each of these three tools can be enabled and disabled on demand. At the moment, our implementation does not support the encoding of additional properties using color or size. Further studies are required to find out how such point encodings affect the human perception when being blended. Additionally, sunspot plots can only be used to visualize single-class scatter plots. In the future, multiple classes could be supported by blending multiple surfaces, each representing a specific class. Similarly, it must be investigated how sunspot plots behave in combination with other visualizations. For example when used as overlay, especially because sunspot plots are already the result of a blending operation.

Throughout the whole paper, we have used the blending function described in Equation 5. We found that this blending function does not only provide a sensible transition between dense and sparse regions, but also results in most readable visualizations. We also experimented with other blending functions, such as curvature or normal-orientation based blending. However, we found that these functions were less predictable, and that the intuitive interpretation of distance-based blending is most advantageous.

When using frequency-based methods, the visual quality of the result is highly dependent on kernel properties such as shape and size. We therefore allow users to dynamically adjust these parameters to change the degree of aggregation of the data samples. This can lead to different visualization results within the same data set. However, this is a drawback of all frequency-based methods. In our current implementation, we therefore provide a preset of kernel properties that were derived through empirical trials.

In our experiments, we found that sunspot plots are scalable in terms of data size and most effective for data with high sample density including global and local density variations. Data sets containing only sparse samples are well represented with conventional scatter plots, whereas dense data containing only global variations can be depicted effectively with frequency-based approaches. The same applies to uniformly sampled data sets since sunspot plots will then not encode additional information.

Although illumination of sunspot plots was not shown to significantly improve the assessment of the density information, there might be other tasks that would benefit of an illumination model which were not evaluated in our user study. In our personal and anecdotal experience, illumination did improve the perception of local density variations, where high-frequency changes in local density tend to manifest themselves as rough vs. smooth surface appearance. However, such differences were not captured in our experiment, as they do not correspond to the most typical tasks in scatter plots, and yet need to be confirmed through further empirical studies covering a wider set of tasks.

Finally, our current GPU-based implementation is capable of rendering several thousands of data samples in real-time. If one were to visualize even larger data sets, for example several millions of sample points, the frame rate would no longer be interactive. To be able to process such large amounts of data, additional preprocessing steps such as subsampling would be required.

9. Conclusion

We have presented sunspot plots, a visualization approach for dense scattered bivariate data. Sunspot plots allow a smooth transition between discrete and continuous data representations based on the local sample density in the image space. As such, sunspot plots alleviate the common drawbacks of discrete and continuous frequency-based visualization approaches for scattered data. The smooth blending between the discrete and continuous representations enables the visualization of main trends in dense areas while still preserving outliers in sparse regions. We have demonstrated that our technique is able to handle various data sets while still allowing for interactive exploration. Finally, we have evaluated the effectiveness of our approach in a user study, indicating that blending discrete and continuous information does not impair the estimation of densities while providing additional information of individual data points.

Acknowledgment

The research presented in this paper was supported by the MetaVis project (#250133) funded by the Research Council of Norway. In addition, we want to thank Michaël Aupetit for providing us with a variety of different scatter plot data sets for our user study.

References

[AASB19] ABBAS M. M., AUPETIT M., SEDLMAIR M., BENSMAIL H.: ClustMe: A Visual Quality Measure for Ranking Monochrome Scatterplots based on Cluster Patterns. *Computer Graphics Forum 38*, 3 (2019), 225–236. doi:10.1111/cgf.13684. 9

[ALG*13] ARLT R., LEUSSU R., GIESE N., MURSULA K., USOSKIN I.: Sunspot positions and sizes for 1825-1867 from the observations by Samuel Heinrich Schwabe. *Monthly Notices of the Royal Astronomical Society 433*, 4 (2013), 3165—-3172. doi:10.1093/mnras/stt961. 2

[ASAB19] AUPETIT M., SEDLMAIR M., ABBAS M. M., BAGGAG A.: Toward Perception-Based Evaluation of Clustering Techniques for Visual Analytics. In *Proc. IEEE Visualization* (2019), pp. 141–145. doi:10.1109/VISUAL.2019.8933620. 9

[BCLC97] BRODBECK D., CHALMERS M., LUNZER A., COTTURE P.: Domesticating Bead: Adapting an Information Visualization System to a Financial Institution. In *Proc. IEEE InfoVis* (1997), pp. 73–80. doi:10.1109/INFVIS.1997.636789. 3

[BGR06] BÜRING T., GERKEN J., REITERER H.: User Interaction with Scatterplots on Small Screens - A Comparative Evaluation of Geometric-Semantic Zoom and Fisheye Distortion. *IEEE Transactions on Visualization and Computer Graphics 12*, 5 (2006), 829–836. doi:10.1109/TVCG.2006.187. 3

[Bou19] BOUTREUX C.: The GeoNames geographical database. http://www.geonames.org/, Accessed: March 2019. 7

[BS06] BERTINI E., SANTUCCI G.: Give Chance a Chance: Modeling Density to Enhance Scatter Plot Quality through Random Data Sampling. *Information Visualization 5*, 2 (2006), 95–110. doi:10.1057/palgrave.ivs.9500122. 3

[BW08] BACHTHALER S., WEISKOPF D.: Continuous Scatterplots. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1428–1435. doi:10.1109/TVCG.2008.119. 4

[BW09] BACHTHALER S., WEISKOPF D.: Efficient and Adaptive Rendering of 2-D Continuous Scatterplots. *Computer Graphics Forum 28*, 3 (2009), 743–750. doi:10.1111/j.1467-8659.2009.01478.x. 4

[CCF95] CARPENDALE M. S. T., COWPERTHWAITE D. J., FRACCHIA F. D.: 3-Dimensional Pliable Surfaces: For the Effective Presentation of Visual Information. In *Proc. ACM UIST* (1995), pp. 217–226. doi:10.1145/215585.215978. 3

[CCKT83] CHAMBERS J. M., CLEVELAND W. S., KLEINER B., TUKEY P. A.: *Graphical Methods for Data Analysis*. Chapman and Hall/Cole Publishing Company, 1983, p. 107. doi:10.1201/9781351072304. 3

[CCM*14] CHEN H., CHEN W., MEI H., LIU Z., ZHOU K., CHEN W., GU W., MA K.-L.: Visual Abstraction and Exploration of Multi-class Scatterplots. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1683–1692. doi:10.1109/TVCG.2014.2346594. 3

[CEJ*18] CHEN H., ENGLE S., JOSHI A., RAGAN E. D., YUKSEL B. F., HARRISON L.: Using Animation to Alleviate Overdraw in Multi-class Scatterplot Matrices. In *Proc. ACM CHI* (2018), pp. 417:1–417:12. doi:10.1145/3173574.3173991. 3

[CGZ*19] CHEN X., GE T., ZHAN J., CHEN B., FU C.-W., DEUSSEN O., WANG Y.: A Recursive Subdivision Technique for Sampling Multi-class Scatterplots. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (2019), 729–738. doi:10.1109/TVCG.2019.2934541. 3

[CLNL87] CARR D., LITTLEFIELD R. J., NICHOLSON W. L., LITTLEFIELD J. S.: Scatterplot Matrix Techniques for Large N. *Journal of the American Statistical Association 82*, 398 (1987), 424–436. doi:10.2307/2289444. 2, 3

[CM84] CLEVELAND W. S., MCGILL R.: The Many Faces of a Scatterplot. *Journal of the American Statistical Association 79*, 338 (1984), 807–822. doi:10.2307/2288711. 2

[Cor19] CORNUT O.: ImGui. https://github.com/ocornut/imgui, Accessed: March 2019. 6

[DE02] DIX A., ELLIS G.: By Chance Enhancing Interaction with Large Data Sets Through Statistical Sampling. In *Proc. Working Conference on Advanced Visual Interfaces* (2002), pp. 167–176. doi:10.1145/1556262.1556289. 3

[DLH11] DAAE LAMPE O., HAUSER H.: Interactive Visualization of Streaming Data with Kernel Density Estimation. In *Proc. IEEE PacificVis* (2011), pp. 171–178. doi:10.1109/PACIFICVIS.2011.5742387. 3

[DP03] DUPONT W. S., PLUMMER W. D. J.: Density Distribution Sunflower Plots. *Journal of Statistical Software 8*, 3 (2003), 1–5. doi:10.18637/jss.v008.i03. 2

[DWA10] DANG T. N., WILKINSON L., ANAND A.: Stacking Graphic Elements to Avoid Over-Plotting. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1044–1052. doi:10.1109/TVCG.2010.197. 2, 4

[EBD05] ELLIS G., BERTINI E., DIX A.: The Sampling Lens: Making Sense of Saturated Visualisations. In *Proc. ACM CHI Extended Abstracts* (2005), pp. 1351–1354. doi:10.1145/1056808.1056914. 3

[ED02] ELLIS G., DIX A.: Density Control Through Random Sampling: an Architectural Perspective. In *Proc. International Conference on Information Visualisation* (2002), pp. 82–90. doi:10.1109/IV.2002.1028760. 3

[ED07] ELLIS G., DIX A.: A Taxonomy of Clutter Reduction for Information Visualisation. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1216–1223. doi:10.1109/TVCG.2007.70535. 3

[FD05] FRIENDLY M., DENIS D.: The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences 41*, 2 (2005), 103–130. doi:10.1002/jhbs.20078. 1

[Gib50] GIBSON J. J.: The Perception of Visual Surfaces. *The American Journal of Psychology 63*, 3 (1950), 367–384. doi:10.2307/1418003. 5

[Har75] HARTIGAN J. A.: Printer graphics for clustering. *Journal of Statistical Computation and Simulation 4*, 3 (1975), 187–213. doi:10.1080/00949657508810123. 2

[HDS*10] HAO M., DAYAL U., SHARMA R., KEIM D. A., JANETZKO H.: Variable Binned Scatter Plots. *Information Visualization 9*, 3 (2010), 194–203. doi:10.1057/ivs.2010.4. 2

[HMS97] HUANG C., MCDONALD J. A., STUETZLE W.: Variable Resolution Bivariate Plots. *Journal of Computational and Graphical Statistics 6*, 4 (1997), 383–396. doi:10.1080/10618600.1997.10474749. 2

[Hun07] HUNTER J. D.: Matplotlib: A 2D graphics environment. *Computing In Science & Engineering 9*, 3 (2007), 90–95. doi:10.1109/MCSE.2007.55. 4, 6, 7

[JHM*13] JANETZKO H., HAO M. C., MITTELSTÄDT S., DAYAL U., KEIM D. A.: Enhancing Scatter Plots Using Ellipsoid Pixel Placement and Shading. In *Proc. Annual Hawaii International Conference on System Sciences* (2013), pp. 1522–1531. doi:10.1109/HICSS.2013.197. 3

[Kei00] KEIM D. A.: Designing Pixel-oriented Visualization Techniques : Theory and Applications. *IEEE Transactions on Visualization and Computer Graphics 6*, 1 (2000), 59–78. doi:10.1109/2945.841121. 3

[KH98] KEIM D. A., HERRMANN A.: The Gridfit Algorithm: An Efficient and Effective Approach to Visualizing Large Amounts of Spatial Data. In *Proc. IEEE Visualization* (1998), pp. 181–188. doi:10.1109/VISUAL.1998.745301. 3

[KHD*09] KEIM D. A., HAO M. C., DAYAL U., JANETZKO H., BAK P.: Generalized Scatter Plots. *Information Visualization 9*, 4 (2009), 301–311. doi:10.1057/ivs.2009.34. 3

[LBH18] LI C., BACIU G., HAN Z.: StreamMap: Smooth Dynamic Visualization of High-Density Streaming Points. *IEEE Transactions on Visualization and Computer Graphics 24*, 3 (2018), 1381–1393. doi:10.1109/TVCG.2017.2668409. 3

[LCCJ20] LECUN Y., CORTES C., CHRISTOPHER J.C. B.: The MINST Database for Handwritten digits. http://yann.lecun.com/exdb/mnist/, Accessed: Februar 2020. 8

[LMv10] LI J., MARTENS J.-B., VAN WIJK J. J.: A Model of Symbol Size Discrimination in Scatterplots. In *Proc. ACM CHI* (2010), pp. 2553–2562. doi:10.1145/1753326.1753714. 2

[LvM09] LI J., VAN WIJK J. J., MARTENS J.-B.: Evaluation of Symbol Contrast in Scatterplots. In *Proc. IEEE PacificVis* (2009), pp. 97–104. doi:10.1109/PACIFICVIS.2009.4906843. 2

[LvM10] LI J., VAN WIJK J. J., MARTENS J.-B.: A Model of Symbol Lightness Discrimination in Sparse Scatterplots. In *Proc. IEEE PacificVis* (2010), pp. 105–112. doi:10.1109/PACIFICVIS.2010.5429604. 2

[MAF15] MATEJKA J., ANDERSON F., FITZMAURICE G.: Dynamic Opacity Optimization for Scatter Plots. In *Proc. ACM CHI* (2015), pp. 2707–2710. doi:10.1145/2702123.2702585. 2

[MG13] MAYORGA A., GLEICHER M.: Splatterplots: Overcoming Overdraw in Scatter Plots. *IEEE Transactions on Visualization and Computer Graphics 19*, 9 (2013), 1526–1538. doi:10.1109/TVCG.2013.65. 2, 5

[Mil20] MILL B.: Boston Marathon Raw Data. https://github.com/llimllib, Accessed: Februar 2020. 7

[MM18] MAHMOOD S., MUELLER K.: An Exploded View Paradigm to Disambiguate Scatterplots. *Computers & Graphics 73* (2018), 37–46. doi:10.1016/j.cag.2018.02.008. 3

[MML12] McGuire M., Mara M., Luebke D.: Scalable Ambient Obscurance. In *Proc. ACM High-Performance Graphics* (2012), pp. 97–103. doi:10.2312/EGGH/HPG12/097-103. 6, 9

[MPOW17] Micallef L., Palmas G., Oulasvirta A., Weinkauf T.: Towards Perceptual Optimization of the Visual Design of Scatterplots. *IEEE Transactions on Visualization and Computer Graphics 23*, 6 (2017), 1588–1599. doi:10.1109/TVCG.2017.2674978. 3

[PD84] Porter T., Duff T.: Compositing Digital Images. *ACM SIGGRAPH Computer Graphic 18*, 3 (1984), 253–259. doi:10.1145/964965.808606. 6

[Pho75] Phong B. T.: Illumination for Computer Generated Pictures. *Communications of the ACM 18*, 6 (1975), 311–317. doi:10.1145/360825.360839. 5, 6, 9

[Pla05] Playfair W.: *The Commercial and Political Atlas and Statistical Breviary*. Cambridge University Press, 2005. doi:10.1162/leon.2007.40.2.202a. 2

[Ram88] Ramachandran V.: Perception of Shape from Shading. *Nature 331* (1988), 163–166. doi:10.1038/331163a0. 3

[RGE19] Raidou R. G., Gröller E., Eisemann M.: Relaxing Dense Scatter Plots with Pixel-Based Mappings. *IEEE Transactions on Visualization and Computer Graphics 25*, 6 (2019), 2205–2216. doi:10.1109/TVCG.2019.2903956. 3

[SBG00] Sprenger T. C., Brunella R., Gross M. H.: H-BLOB: A Hierarchical Visual Clustering Method Using Implicit Surface. In *Proc. IEEE Visualization* (2000), pp. 61–68. doi:10.1109/VISUAL.2000.885677. 4

[SFB94] Stone M. C., Fishkin K., Bier E. A.: The Movable Filter As a User Interface Tool. In *Proc. ACM CHI* (1994), pp. 306–312. doi:10.1145/191666.191774. 3

[SG18] Sarikaya A., Gleicher M.: Scatterplots: Tasks, Data, and Designs. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2018), 402–412. doi:10.1109/TVCG.2017.2744184. 2, 4, 8, 9, 10

[SGG16] Staib J., Grottel S., Gumhold S.: Enhancing Scatterplots with Multi-Dimensional Focal Blur. *Computer Graphics Forum 35*, 3 (2016), 11–20. doi:10.1111/cgf.12877. 4

[Sil86] Silverman B. W.: *Density Estimation for Statistics and Data Analysis*. Chapman and Hall/CRC, 1986. doi:10.1201/9781315140919. 4

[TGC94] Trutschl M., Grinstein G. G., Cvek U.: Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. In *Proc. ACM CHI* (1994), pp. 313–317. doi:10.1145/191666.191775. 3

[TGC03] Trutschl M., Grinstein G. G., Cvek U.: Intelligently Resolving Point Occlusion. *IEEE Symposium on Information Visualization* (2003), 131–136. doi:10.1109/INFVIS.2003.1249018. 3

[TGK*17] Tominski C., Gladisch S., Kister U., Dachselt R., Schumann H.: Interactive Lenses for Visualization: An Extended Survey. *Computer Graphics Forum 36*, 6 (2017), 173–200. doi:10.1111/cgf.12871. 3

[The06] Theus M.: *Graphics of Large Data Sets: Visualizing a Million - Scaling Up Graphics*. Springer, 2006, pp. 55–72. doi:10.1007/0-387-37977-0_3. 2

[TM19] Thomson G., Macpherson F.: Adelson's checker-shadow illusion. https://www.illusionsindex.org/ir/checkershadow, Accessed: December 2019. 6

[Tru81] Trumbo B. E.: A Theory for Coloring Bivariate Statistical Maps. *The American Statistician 35*, 4 (1981), 220–226. doi:10.2307/2683294. 2

[TSW*07] Tory M., Sprague D., Wu F., So W. Y., Munzner T.: Spatialization Design: Comparing Points and Landscapes. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1262–1269. doi:10.1109/TVCG.2007.70596. 5

[vd03] van Liere R., de Leeuw W.: GraphSplatting: Visualizing Graphs as Continuous Fields. *IEEE Transactions on Visualization and Computer Graphics 9*, 2 (2003), 206–212. doi:10.1109/TVCG.2003.1196007. 3

[vT01] van Wijk J. J., Telea A.: Enridged Contour Maps. In *Proc. IEEE Visualization* (2001), pp. 69–543. doi:10.1109/VISUAL.2001.964495. 4

[WB04] Waldeck C., Balfanz D.: Mobile Liquid 2D Scatter Space (ML2DSS). In *Proc. International Conference on Information Visualisation* (2004), pp. 494–498. doi:10.1109/IV.2004.1320190. 2

[Wil05] Wilkinson L.: *The Grammar of Graphics*. Springer, 2005. doi:10.1007/0-387-28695-0. 2

[WLS98] Woodruff A., Landay J., Stonebraker M.: Constant Density Visualizations of Non-uniform Distributions of Data. In *Proc. ACM UIST* (1998), pp. 19–28. doi:10.1145/288392.288397. 2

[WVBT16] Wambecke J., Vergne R., Bonneau G.-P., Thollot J.: Automatic lighting design from photographic rules. In *Proc. Eurographics Workshop on Intelligent Cinematography and Editing* (2016), pp. 1–8. doi:10.2312/wiced.20161094. 6

[ZBDS12] Zinsmaier M., Brandes U., Deussen O., Strobelt H.: Interactive Level-of-Detail Rendering of Large Graphs. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2486–2495. doi:10.1109/TVCG.2012.238. 3

[ZIK98] Zhukov S., Iones A., Kronin G.: An ambient light illumination model. In *Proc. Eurographics Workshop on Rendering* (1998), pp. 45–55. doi:10.1007/978-3-7091-6453-2_5. 6