# Relaxing Dense Scatter Plots with Pixel-Based Mappings

Renata G. Raidou, M. Eduard Gröller, and Martin Eisemann

**Abstract—**Scatter plots are the most commonly employed technique for the visualization of bivariate data. Despite their versatility and expressiveness in showing data aspects, such as clusters, correlations, and outliers, scatter plots face a main problem. For large and dense data, the representation suffers from clutter due to overplotting. This is often partially solved with the use of density plots. Yet, data overlap may occur in certain regions of a scatter or density plot, while other regions may be partially, or even completely empty. Adequate pixel-based techniques can be employed for effectively filling the plotting space, giving an additional notion of the numerosity of data motifs or clusters. We propose the Pixel-Relaxed Scatter Plots, a new and simple variant, to improve the display of dense scatter plots, using pixel-based, space-filling mappings. Our Pixel-Relaxed Scatter Plots make better use of the plotting canvas, while avoiding data overplotting, and optimizing space coverage and insight in the presence and size of data motifs. We have employed different methods to map scatter plot points to pixels and to visually present this mapping. We demonstrate our approach on several synthetic and realistic datasets, and we discuss the suitability of our technique for different tasks. Our conducted user evaluation shows that our Pixel-Relaxed Scatter Plots can be a useful enhancement to traditional scatter plots.

**Index Terms—**Scatter plots, overplotting, pixel-based technique, space-filling technique

## 1 Introduction

Scatter plots are common in visualizing bivariate data [1]. The traditional scatter plot is defined by a space of two continuous, orthogonal dimensions used for the depiction of data points. Scatter plots are simple and intuitive—yet, powerful and versatile—representations, which have the ability of communicating significant aspects of the data, such as clusters, correlations, or outliers. Their advantages have led them to be vastly employed in a variety of exploratory and presentation tasks [2], [3]. However, with an increasing number of data points, they become less effective, due to data overplotting.

*Overplotting*—also called overdrawing—describes a situation where two or more data points overlap, and their marking points are drawn on top of each other [2]. As the number of plotted data points grows, the available plotting space decreases, and the number of overlapping points increases. Therefore, it often becomes difficult to perceive every single data point within the dataset, and subsequently to explore and identify patterns, relations, or outliers in the data—in particular, regarding their density or numerosity. Several approaches have been proposed in the past, to address overplotting in scatter plots. For example, transparency is often employed when rendering data points, to facilitate the identification of high density or high overlap areas [4]. Variants of density plots and contours have also been used, to the same end [2]. We will review even more advanced approaches and their benefits and limitations in detail, in Section 2.

An additional issue with scatter plots is that they may come with a lot of blank plotting space—depending on the represented dataset. In essence, the representation is not optimizing the *use of the available screen space*. Some plotting regions are overused, while others are unexploited. Pixel-based techniques that effectively use the plotting space could be a solution to this problem, but have not been investigated for scatter plots, yet. The main reason is that pixel-based, space-filling techniques displace data points to fit the new space, which might compromise clarity and readability.

In this work, we propose the *Pixel-Relaxed Scatter Plot (PRSP)*, a new and simple variant, which can be used in addition to the traditional scatter plots. It uses pixel-based, space-filling mappings, in order to improve the display of otherwise cluttered and dense scatter plots, enhancing density and numerosity data information, in particular. The *contribution* of our work is the introduction of the concept of a PRSP as an extension that makes better use of the plotting canvas, avoiding data overplotting and optimizing space coverage. PRSPs are a useful supplement to the traditional scatter plots, suitable for completing several tasks, such as cluster detection, or density and numerosity estimation.

## 2 Related Work

Overplotting is a common phenomenon that may hide important information about the plotted data. Reducing

- *R.G. Raidou is with the Institute of Visual Computing & Human-Centered Technology, TU Wien, Vienna 1040, Austria. E-mail: rraidou@cg.tuwien.ac.at.*
- *M. Eduard Gröller is with the Institute of Visual Computing & Human-Centered Technology, TU Wien, Vienna 1040, Austria, and also with the the VRVis Research Center, Wien 1220, Austria.*
- *M. Eisemann is with the TH Köln, Köln50968, Germany. E-mail: martin.eisemann@th-koeln.de.*

*Manuscript received 5 Oct. 2018; revised 21 Dec. 2018; accepted 4 Jan. 2019. Date of publication 15 Mar. 2019; date of current version 1 May 2019. (Corresponding author: Renata G. Raidou.) Recommended for acceptance by R. Maciejewski, J. Seo, and R. Westermann. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TVCG.2019.2903956*

overplotting is a widely researched topic in Information Visualization—not only with respect to scatter plots. It relates to the general topic of quality and clutter reduction within Information Visualization representations [3]. According to the taxonomy proposed by Ellis and Dix [5], clutter reduction techniques are classified into techniques that affect the *appearance* of the representation, techniques that involve *spatial distortion*, and *animation*. In this section, we focus on the first two categories, omitting animation.

*Techniques Affecting the Appearance of the Representation*. This category includes *sampling*, i.e., the random selection of a subset of the data points [6], [7], [8], [9], and *filtering* [10], [11], [12], i.e., the selection of a subset of the available data points that are interesting or fulfill specific requirements. Both filtering and sampling, although straightforward, often come with a loss of information. This category also includes methods based on *clustering* of data points [13], which entail the limitation that the entire method depends on the complexity and accuracy of the employed algorithm. Changes in the *appearance of the data points* [14], such as changes in the point size [15], [16], [17], opacity [15], [18], [19], or focal blur [15], [20], [21] fall also within this category. Still, for most, overplotting persists with a highly increasing number of points. In the work of Micallef et al. [4], the authors adapt the visual design of scatter plots based on a cost function that combines aspects of the human visual system and data aspects. Yet, the cost function requires a one-time careful balancing of its terms by the designer.

*Density plots* can also be considered as an enhancement that affects the appearance of scatter plots, whether these take the form of grey-scale, colored, or kernel density estimation plots [22], [23], [24], [25], contours [22], [25], [26], [27], hexagon bins [22], [25], or variable bins [28]. All these techniques aggregate discrete data points and do not plot them distinctively, dealing effectively with overdrawing and reducing clutter in the representation. However, they are often based on sensitive statistical models, such as kernel density estimators [29], and do not show the original data points. This might affect the accuracy and expressiveness of the representation. In particular, contouring or binning are not good for dense plots, as they create perceptually non-smooth regions that become difficult to compare visually [24]. Visualizations using kernel density estimation can be considered to be similar to a low pass filtering of the scatter plot data. Hence, the result of such approaches might not be necessarily unique.*Techniques Involving Spatial Distortion of the Representation*. This category covers techniques for *displacement* [15], [30], [31]—even in the third dimension [27], [32]. Still, displacement has to be used with care, as it might interfere with the perception of patterns. Other techniques in this category employ *distortion* [33], [34], [35], including approaches, such as zooming or Fish-eye lenses [36]. Closer to our work, space-filling and pixel-plotting approaches have also been proposed and can be included in the spatial distortion category. *Space filling* can be described as a non-overlapping way of rearranging representations in the screen space [2], [5], [30], [37], [38]. This approach has been applied on treemaps and sunburst visualizations, but there is some previous work also in the domain of scatter plots, such as the generalized scatter plots of Keim et al. [39] and the continuous scatter plots of Bachthaler et al. [40]. They both take advantage of empty plotting space and combine it with the benefits of density estimation.

*Pixel plotting* refers to a non-overlapping mapping of each data point to a single pixel, in order to make better use of the available screen space [5], [41], [42]. The main benefit of this approach is that the representations can achieve a high resolution, which is practically the same as the screen resolution. Some of the most well-known examples of pixel-based techniques include the previous work of Keim et al. [43] that follows a recursive pattern to visualize large numbers of data points in pixel-based arrangements. This has been successfully applied in bar charts to create pixel-based extensions of the representation, which facilitate the visualization of large amounts of data [44]. Concerning scatter plots, Fekete and Plaisant [45] describe interactive pixel-based techniques to handle a million data points, so that they remain visible and manageable.

## 3 The Pixel-Relaxed Scatter Plot (PRSP)

*Requirements*. Scatter plots are used for a variety of tasks, which have been classified by Sarikaya et al. [2] and Behrisch et al. [3]. We will discuss in Section 4 which of the tasks from the taxonomy of Sarikaya et al. [2] are tackled in our work. In our case, we do not intend to replace scatter plots, but to provide a supplemental variant that satisfies criteria for clutter reduction techniques, as described by Ellis and Dix [5]. We focus on *avoiding overlap* in the display of data points, avoiding loss of information. Additionally, we take care of retaining the *identification of the overlap density*, i.e., the amount of overplotting and the numerosity of data points in specific motifs. *Scalability* to large datasets is also crucial for an increasing number of data points. Moreover, we would like to *preserve as much as possible the spatial information* of the data points, but in case distortion cannot be avoided, it should at least be conveyed to the user. Finally, it is necessary for our variant to *link back to the original representation*. Given the previous requirements, we aim at obtaining a result that does not obstruct the detection of patterns and trends in the data—in particular, clusters. Additionally, we aim at retaining the ability to compare features in the data, such as correlations across, or numerosity and density within scatter plots. In the next sections, we discuss how and to which extent our proposed Pixel-Relaxed Scatter Plots (PRSPs) satisfy these points.*Overview of the Steps of our Proposed Technique*. The design of our proposed PRSP consists of three main steps. First, the scatter plot data points are *mapped to unique image pixels*, in a way that the previously mentioned criteria are satisfied. Second, the *relaxation* of the data points, i.e., the introduced distortion of the spatial information, is measured, encoded, and communicated to the intended user. Third, suitable *visual encodings* that ensure linking to the original scatter plot representation are discussed. In the following sections, we describe alternatives for each step.

### 3.1 Point-to-Pixel Mapping

The first step of the PRSP generation matches each data point from the original scatter plot to a pixel in the output image, as presented in Fig. 1. Before the actual mapping, we conduct a preliminary step for the computation of the required resolution in pixels of the resulting *PRSP canvas*. For a scatter plot of $N$ data points, each data point has to be mapped to a unique pixel on the PRSP canvas. The PRSP canvas consists of

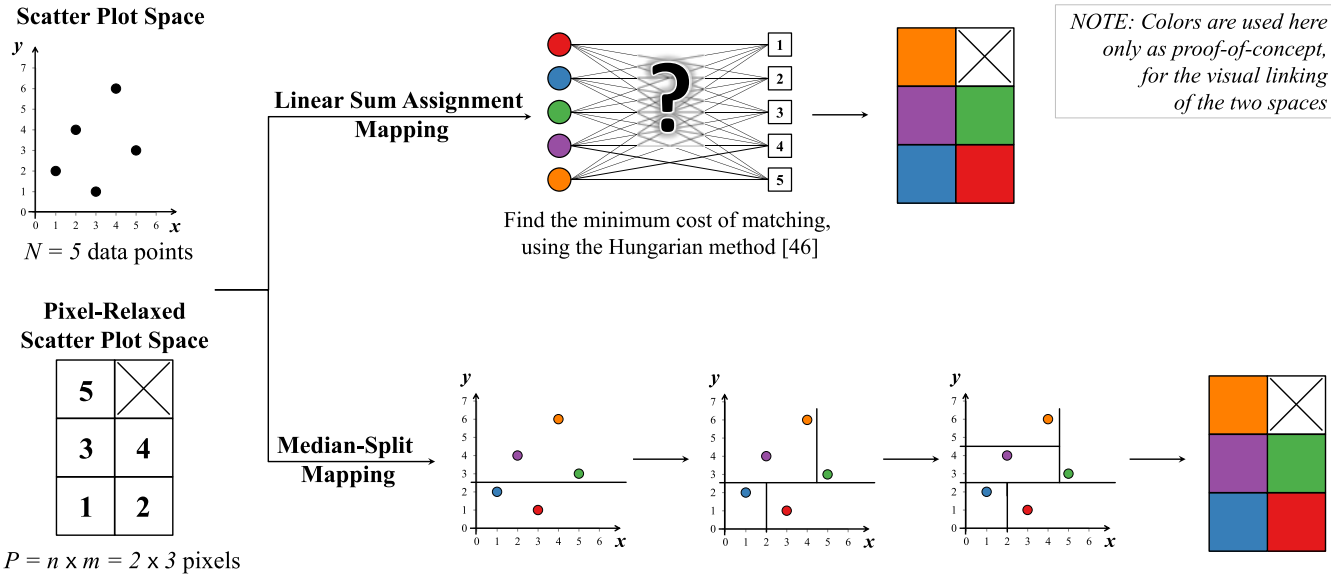


Fig. 1. Schematic depiction of the two different approaches employed for the point-to-pixel mapping (Linear Sum Assignment and Median-Split Mapping). This example uses a toy dataset with five data points for demonstration purposes, only.

$P = n \times m$ pixels, where $n = \lfloor \sqrt{N} \rfloor$ is the width of the resulting canvas and $m = \lceil N/n \rceil$ the height. In this way, we account also for non-square numbers of data points and we always have $P \geq N$. If $P > N$, the excess pixels will be left unused (white) in the final image representation. Other aspect ratios could be employed, but we prefer the $1:1$ aspect, being the most common and simplest. For example, in Fig. 1, we depict a simple example with $N = 5$ data points. In this case, we construct a canvas of $2 \times 3$ pixels, where the first 5 pixels are filled and the last is not. This is also shown with the intentionally-left white pixels, at the right upper corners of the examples in Figs. 4 and 5.

After the creation of the PRSP canvas, we determine an injective function for the mapping of the data points to the canvas pixels. This unique mapping will ensure that our data points avoid overlapping onto each other. We want to distort the original scatter plot space as little as possible, to preserve the identification of data patterns and to convey the amount of overlap. To this end, we developed two different point-to-pixel mapping techniques, with different strengths and limitations, as described in the following sections. To simplify our explanation, and without loss of generality, we assume in the following subsections that all datasets are bivariate. We denote the horizontal and vertical dimension $x$ and $y$, respectively. Furthermore, we normalize each data point $d_i,\ i \in 1 \ldots N$ in the $x$ and $y$ direction, to lie within the range of the output PRSP canvas $[0, n-1] \times [0, m-1]$.

### 3.1.1 Linear Sum Assignment Point-to-Pixel Mapping

To minimize distortion of the spatial information and to preserve as much as possible the coherence of the original scatter plot, we pose the point-to-pixel mapping as an optimization problem. Our goal is to minimize the pixel displacement of each data point from its position in the scatter plot to the position in the PRSP. This can be formulated as a *linear sum assignment problem*, which can be summarized by the function $\min \sum_i \sum_j C_{i,j} X_{i,j}$, $X$ being a boolean matrix, where $X_{i,j} = 1$, if and only if data point $d_i$, with $i \in 1 \ldots N$ is assigned to pixel $p_j$, with $j \in 1 \ldots P$. Otherwise, $X_{i,j} = 0$. We denote as $C$ the associated cost matrix of assigning data point $d_i$ to pixel $p_j$. Subsequently, we create a second dataset using the pixel position of $N$ pixels in the output canvas. For simplicity reasons, if $N \neq P$, we remove pixels from the last column until we have $N$ pixel positions. Given these two datasets, we fill the cost matrix $C$ with the euclidean distance between each pair of the data point $d_i$ and the potential pixel position $p_j$ in the output. To this end, we employ the Hungarian Method [46] to compute the optimal assignment with minimal distortion with reference to the euclidean distance. However, this algorithm comes with a complexity of $\mathcal{O}(N^3)$, which makes it slow for a high number of data points.

### 3.1.2 Median-Split Point-to-Pixel Mapping

The median-split point-to-pixel mapping is a recursive mapping method used as a fast, approximate alternative to the linear sum assignment mapping. This mapping builds a left-balanced tree that assigns data points to pixels. Starting with the entire dataset and all pixels in the canvas, the algorithm first sets the $y$-axis to be the splitting axis. Then, it sorts the data points and pixels along this axis. Subsequently, we calculate the median pixel value along the splitting axis. This value determines a splitting line, orthogonal to the splitting axis. All $k$ pixels with a position smaller than the splitting axis are assigned to the left child-node. Similarly, the first $k$ data points are also assigned to the left child-node. Respectively, all other pixels and data points are assigned to the right child-node. The algorithm continues recursively, until we have a unique, one-to-one mapping of data points to pixels. A rudimentary example of the median-split is presented on the bottom half of Fig. 1. As we mentioned earlier, in the case of $P > N$, some pixels will not be mapped. These pixels are simply left empty, but gather in one corner of the image due to the left-balanced assignment. This approach preserves most of the locality of the data points in the resulting image, while being efficient to compute in $\mathcal{O}(N \log N)$, as opposed to $\mathcal{O}(N^3)$ of the linear sum assignment.

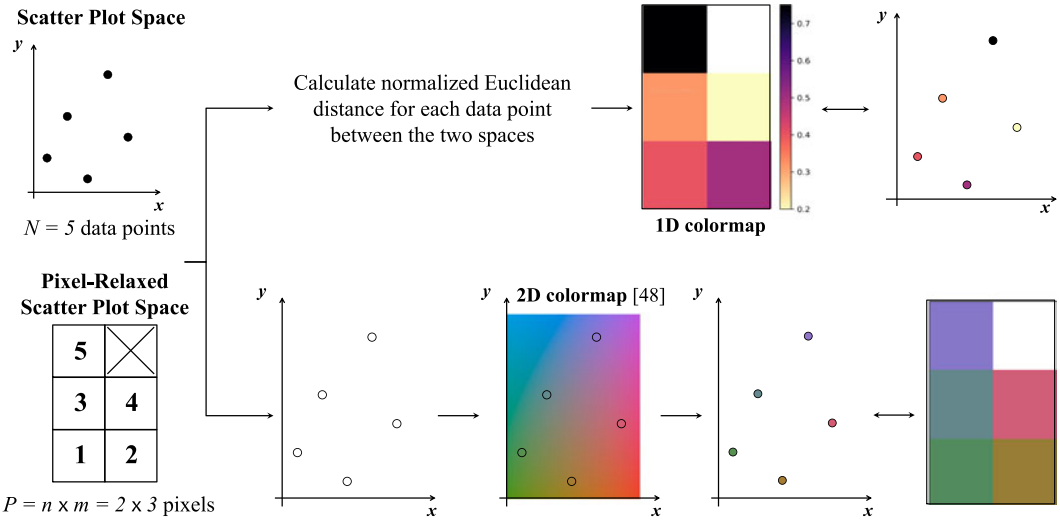


Fig. 2. Process for the calculation, encoding and visualization of the relaxation of the pixels in the PRSP representation, performed in two ways.

### 3.2 Relaxation Mapping and Encoding

The point-to-pixel mapping techniques *relax*, i.e., displace the data points from their original position in the scatter plot to a position on the PRSP canvas. The second part of our approach requires the calculation and encoding of this introduced displacement of the data points. We propose two alternatives to convey the relaxation in the PRSP in the following sections.

#### 3.2.1 Using Normalized Euclidean Distance

The first approach calculates the relaxation of data points from their original position in the scatter plot to the pixel position in the PRSP, based on the euclidean distance. Let $d_i,\ i \in 1 \dots N$ be the normalized position of the $i$th data item on the PRPS canvas, in the range $[0, n-1] \times [0, m-1]$, as described at the beginning of Section 3.1. Let $p_i,\ i \in 1 \dots N$ be the pixel position assigned to $d_i$ from the point-to-pixel mapping. The relaxation value is given by the euclidean distance, as $||d_i - p_i||$. The calculated relaxation value of each canvas pixel is color encoded to indicate the amount of displacement of each pixel, relatively to the original position within the scatter plot. We employ a perceptually uniform 1D colormap for the encoding of the relaxation, as shown in the upper part of Fig. 2—in particular, the inverted `magma` colormap from the `matplotlib` library, which covers a wide perceptual range in brightness. The brighter segment of the colormap draws attention to the non-displaced parts of the scatter plot, preserving the motifs in the data and giving a notion of "peaks and valleys" in the images.

#### 3.2.2 Using 2D Colormaps

A second approach to encode the relaxation of the data points with respect to their position within the PRSP requires to color encode the original data points. In this approach, we overlay a 2D colormap over the original scatter plot space. Then, we assign to each data point the color at its scatter plot position. The assigned color is, subsequently, propagated to the respective pixel in the resulting PRSP. In this way, the spatial information of the data points from the original scatter plot is propagated to the pixels of the PRSP. Different colormaps have been considered [47]. The 2D colormap of Bremm et al. [48] was identified as the most appropriate one, due to its localization and identification properties, as discussed in the paper of Bernard et al. [47]. Still, the use of a 2D colormap requires the presence of a legend. The use of 2D colormaps for the encoding has two purposes. First, it represents the amount of displacement of each data point from its original respective position within the 2D colormap—and, consequently, the original scatter plot space. Second, it serves as a link between the scatter plot and the PRSP domain, enabling the discriminability, localization, and traceability of data points. This is particularly important, given that our PRSPs are an additional extension, not a replacement, of scatter plots. An example of the use of 2D colormaps for the encoding of relaxation is depicted in the lower part of Fig. 2.

### 3.3 Visual Linking of the Two Spaces

From the proposed point-to-pixel mappings, the linear sum assignment optimizes for minimal distortion of the space, but still does not avoid it entirely. Also, linking of the PRSP variant to the original scatter plot representation is required. To show the distortion and to link PRSPs to scatter plots, we employ several visual encodings. In Section 3.2, we introduced two visual encodings to either encode the amount of displacement, or to create a visual link between the PRSP and the position of the data points in the scatter plot. In combination with the 2D colormap, we can employ additional encodings in an enlarged PRSP to show explicitly the relaxation. For example, we render a *circular glyph* at the center of each pixel, where the relaxation—measured as the euclidean distance between the original scatter plot data points and the PRSP pixels—is encoded in several ways, as inspired by the guidelines of Borgo et al. [49]. The relaxation is mapped to the opacity or area of the circular glyphs. Other encodings can also be used, such as the density of *hatching*—or partial hatching—on top of each pixel, or *simple lines* indicating the "flow" of data points from the scatter plot space to the pixel positions in the PRSP canvas. The lines can also be drawn as *arrows*, but the additional arrow tips add clutter to the view. However, some initial tests with an increasing number of data points indicated that circular glyphs, hatching and arrows are not performing

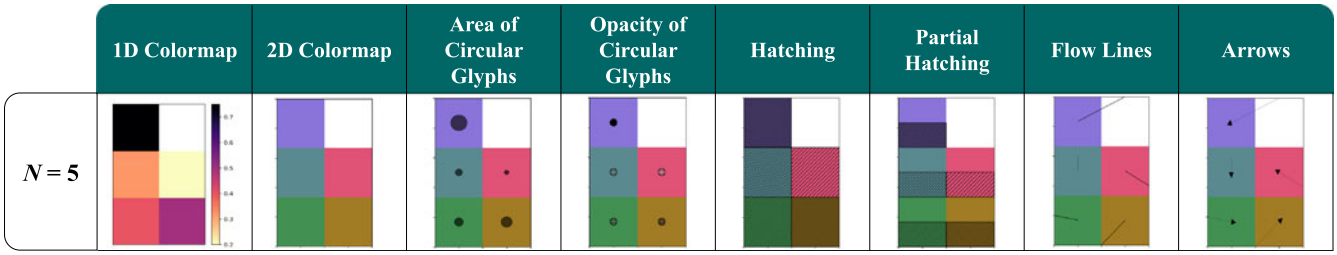


Fig. 3. Visual encodings, employed for the pixel relaxation in PRSPs and for obtaining a spatial link to the data points of the original scatter plot.

well, as they either become small, or clutter the view. Among the investigated alternatives, depicted in Fig. 3, we stick to the flow lines, as an intuitive metaphor for displacement.

## 4 Results

In this section, we present the results of using the PRSP on two different kinds of datasets. First, inspired by previous work [2], [24], we tested the behavior, task support and performance of the PRSPs on a number of synthetic cases with two-dimensional data, containing predefined patterns and structures. Second, to demonstrate a realistic usage scenario of our approach, we use some well-established datasets from the XmdvTool [50], and other datasets, more realistic in size and appearance than the synthetic stimuli. We created the latter using the PCDC tool [51] for the purposes of this work. In the final part of this section, we show how to extend PRSPs for data with more than two dimensions within a PRSP matrix, similar to a scatter plot matrix. With the examples of the upcoming sections, we intend to provide a deeper understanding into PRSPs, as well as their advantages and limitations.

### 4.1 Results with Synthetic Stimuli Data

The recent work of Sarikaya et al. [2] provided a taxonomy of abstracted tasks that can be performed with scatter plots, as well as some basic synthetic stimuli, i.e., sample distributions, which have been used at large in previous literature. Fig. 4 shows our approach applied to these synthetic stimuli, together with their corresponding scatter plots and density plots. Although the employed stimuli have a limited number of data points ($N = 300$), we can already observe some initial interesting facts.

The PRSP is a supplementary representation of the data, where data point overlap is avoided. The linear sum assignment mapping seems to perform better than the median-split mapping, in terms of discernibility and readability of patterns, as well as for the minimization of introduced distortion, which is its intended purpose. In the datasets of Fig. 4, the linear trend, the three clusters and the manifold motif in the data are preserved and clearly depicted in the resulting PRSP. This becomes more obvious in the relaxation encoding with the 1D colormap, but not so clearly with the 2D colormap encoding—with the exception of the linear trend dataset when the flow lines are used. In the dataset with the clusters, the linear

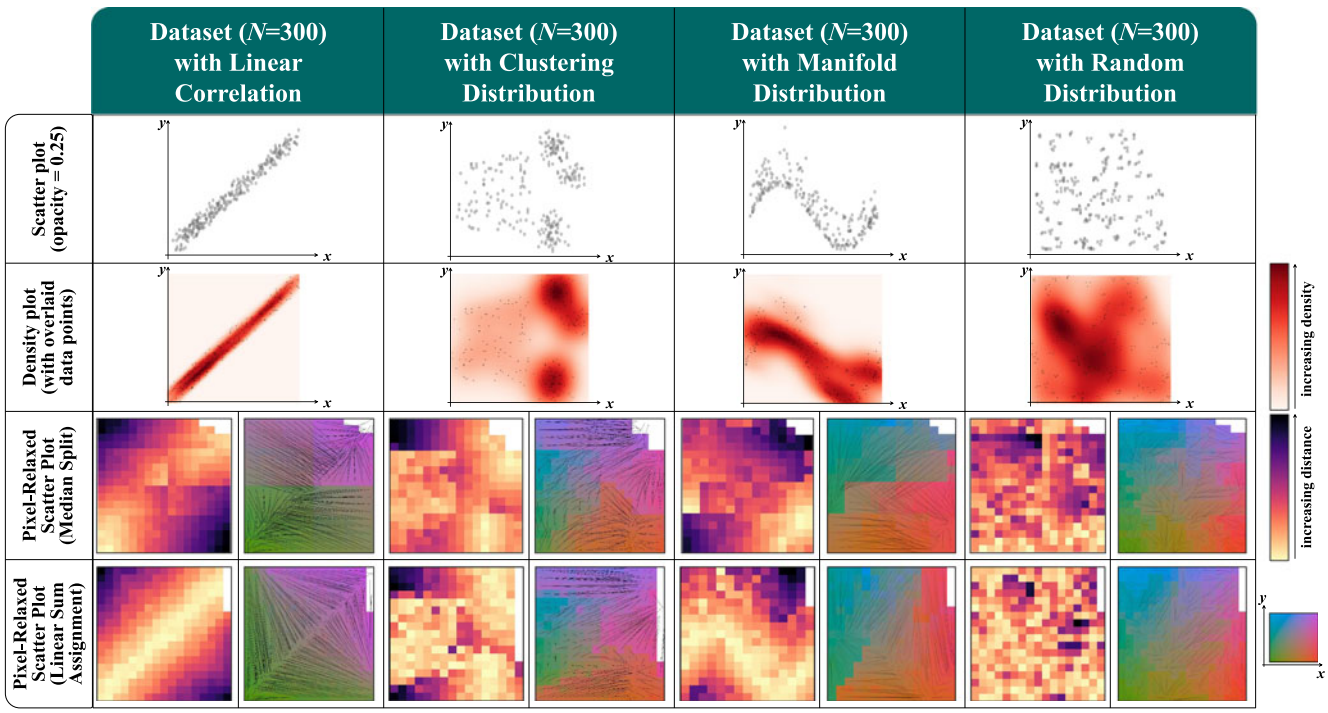


Fig. 4. Results from the application of our approach to four synthetic stimuli, previously proposed by Sarikaya et al. [2]. For each one of the synthetic datasets presented in this figure, we depict the original scatter plot (with reduced opacity), a density plot with the data points overlaid and the resulting Pixel-Relaxed Scatter Plots, computed with the use of the two proposed point-to-pixel mappings and the two relaxation mappings.

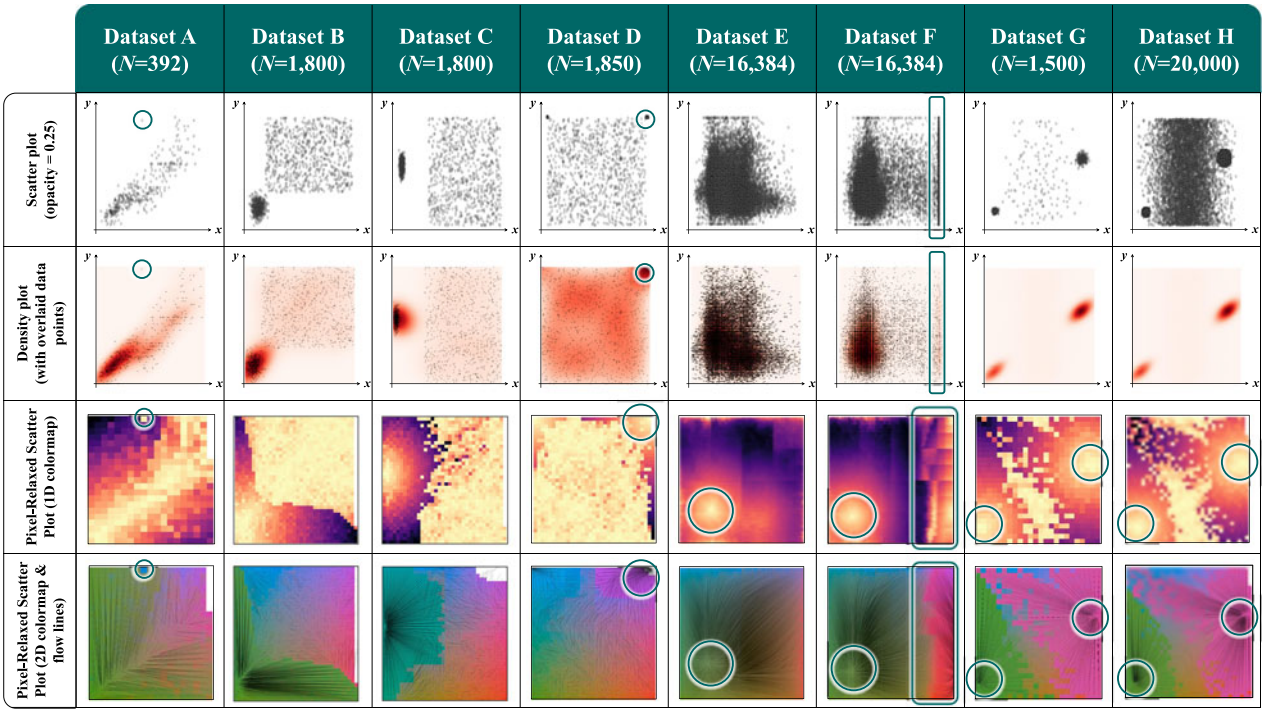


Fig. 5. Results from the application of our approach to eight data sets, obtained from the XmdvTool [50] or created with the PCDC tool [51] for the purposes of this work. For each of the datasets (A-H), we show the original scatter plot (with reduced opacity), a density plot with the original data points overlaid, and our Pixel-Relaxed Scatter Plots, computed with the proposed Linear Sum Assignment Point-to-Pixel Mapping and two relaxation mappings (1D colormap and 2D colormap combined with flow lines).

sum assignment mapping and the median-split mapping are able to convey the presence of the three clusters. The former can additionally give a visual indication of the density and the numerosity of data points, i.e., how spread or tight data points are positioned with respect to each other, and how much larger, with respect to the number of data points, some clusters are. For the cluster preservation, the 2D colormap encoding is particularly helpful. The clusters are depicted as compact regions with similar colors, which are separated with a strong color edge, if located apart from each other. Although the median-split mapping seems to perform well in most cases, it can reduce readability compared to the linear sum assignment, e.g., in the dataset with the manifold. Potentially, PRSPs can be adequate for the exploration of the level of correlation in the represented data, as shown in the dataset with the linear trend, in contrast to the random dataset.

To sum up, we have a first hypothesis that the PRSP variant can be useful for *searching for particular known motifs* in the data, for *exploring the depicted datasets*, for *comparing the density and numerosity* in different regions of the representation and, possibly, for *characterizing distributions* and *determining the level of correlation* in datasets. Yet, it is not possible to understand spatialization of the data due to the introduced distortions and it is not always possible to read the motifs or patterns in all PRSPs.

### 4.2 Results with Realistic Data

In this section, we provide results of the application of our PRSPs on more realistic datasets—containing a higher number of data points and/or more realistic patterns. The obtained results are depicted in Figs. 5, 6 and 7, showing the behavior, characteristics and task support of our introduced enhancement.

Fig. 5 comprises eight datasets. Dataset A contains two of the dimensions of the `Cars` dataset ($x$: horsepower, $y$: displacement) [50], and has $N = 392$ data points. Datasets B and C have been artificially created in the PCDC tool [51] for the purposes of this work. Each of them contains $N = 1,800$ data points, out of which 600 are positioned in the leftmost cluster and the remaining 1200 in the rightmost cluster. Dataset D has also been created by us in the PCDC tool [51], and contains $N = 1,850$ data points, which are strongly overlapping at the right top corner of the representation. Dataset E and F contain two of the dimensions of the `out5d` dataset ($x_E$: potassium, $y_E$: magnetics, $x_F$: thorium, $y_F$: magnetics) [50], and have $N = 16,834$ data points each. Datasets G and H have also

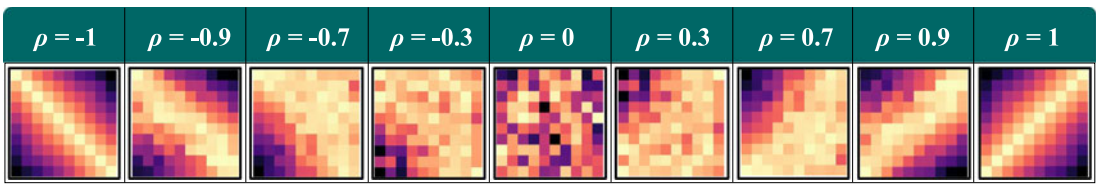


Fig. 6. Results from the application of the PRSPs to nine synthetic stimuli, representing levels of correlation ($\rho$) ranging from $-1$ to 1. The PRSPs have been computed with the Linear Sum Assignment mapping and the relaxation is encoded with the 1D inverted `magma` colormap.

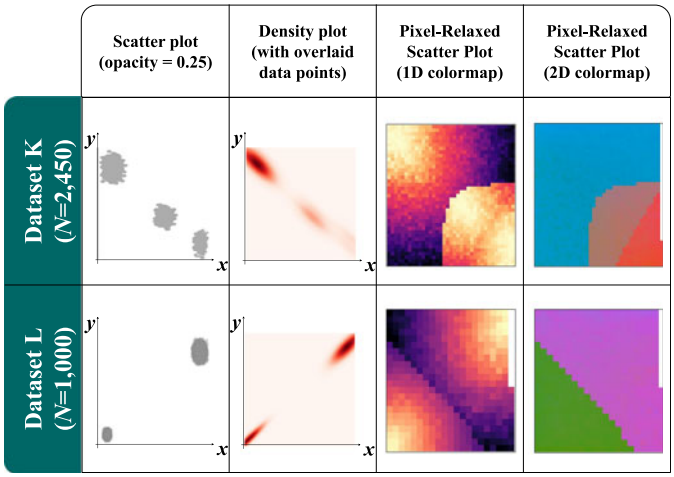


Fig. 7. Results from the application of the PRSPs to two datasets generated with the PCDC tool [51], showing the suitability of our variant for depicting the numerosity and density of clusters.

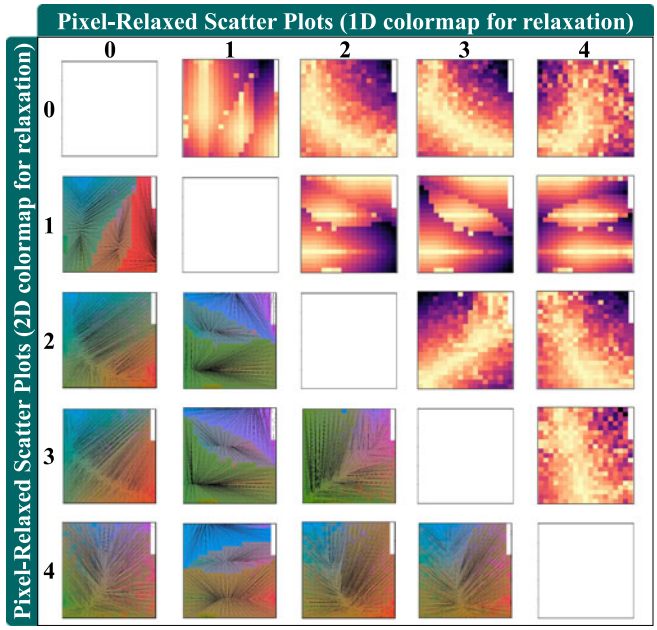


Fig. 8. Results from the application of our approach to the `Cars` dataset [50], in a SPLOM configuration. Above the diagonal, we position the PRSPs with the 1D colormap encoding. Below the diagonal, we position the PRSPs with the 2D colormap encoding, with the additional flow lines. The representations have been computed with the Linear Sum Assignment Point-to-Pixel mapping.

been created in the PCDC tool [51]. The former contains $N = 1,500$ data points and the latter $N = 20,000$, with two clusters and random data points in between to simulate noise in both. All PRSPs in Fig. 5 have been computed using the linear sum assignment method for the point-to-pixel mapping.

For dataset A, we observed two interesting behaviors of our approach. First of all, the correlation that the dataset exhibits in the scatter plot domain is preserved also in the PRSP. This can be seen both in the 1D colormap relaxation encoding, and in the combined 2D colormap and flow lines representation. Second, a strong outlier (annotated) with a high value in the $y$ dimension is also preserved and is discernible in the PRSP. In the case of dataset B, where the number of data points is increasing drastically, we see that the cluster at the bottom-left of the scatter plot is expanded to the empty spaces of the plot, filling the entire PRSP canvas. Also here, the two clusters are discernible, while their numerosity (number of pixels) and density (amount of spread in PRSP space) is also visible. Dataset C also contains two clusters with differing numbers of data points. However, neither in the scatter plot, nor in the density plot, the proportion between the number of data points of the two clusters is visible. Actually, these two can even mislead the user in determining the size of the clusters. On the contrary, our PRSP is able to convey this information—especially, when the 2D colormap is providing a link to the original scatter plot.

For dataset D, the 1D colormap encoded PRSP indicates that most of the data points follow a random pattern. Other data points are concentrated at the right top corner of the representation (annotated), which is visible with the smooth color gradient and the sharp edge structures. Reading these motifs within the 1D colormap encoded PRSP requires some effort and a degree of familiarization. Yet, the 2D colormap alternative, in combination with the flow lines can give us a clearer indication of the structure of the data, indicated by the concentration of the flow lines in this position. For datasets E and F, which contain almost 10 times more data points than the previous three examples, we see that the PRSP representation scales conveniently to larger data sets. In addition to that, information about the overlap density is also conveyed to the user. Both datasets are highly dense at the bottom left of the representation, while dataset F shows also high density for high values of $x$ (annotated). The patterns are preserved and can be identified in the PRSP. The flow lines create easily identifiable and intuitive patterns, which become more informative and aesthetically pleasing with an increasing number of data points. For datasets G and H, we see that the introduction of additional noise in dataset H does not disturb the pattern preservation in the PRSP plots. The two clusters (annotated) are visible with the light yellow color in the 1D colormap PRSP or with the flow lines in the 2D case.

From these examples, we can build the hypothesis that our variant indeed *avoids overlap*, while it also allows to *identify already known motifs*, to *compare the density and numerosity* of different motifs and, possibly, to *explore* and *characterize distributions* after familiarization with the variant and the data spatialization of the introduced distortion. With the examples of Figs. 6 and 7, we can also hypothesize the possibility of *detecting the level of correlation* within a dataset, and the suitability of the PRSPs as a way of representing *density and numerosity of patterns*.

### 4.3 Extending to More Than Two Dimensions

The benefits of our approach can be of particular significance with an increasing number of dimensions, as well. For example, the representation of datasets with more than two dimensions, using the so-called Scatter Plot Matrices (SPLOMs) [52] or GPLOMs [53], reduces significantly the screen space, available for plotting each pair of dimensions. Our PRSPs, if used instead of the traditional scatter plots, can visualize the data in a more readable and aggregative manner, exploiting better the available screen space. An example of our approach is presented in Fig. 8 for the `Cars` dataset [50]. In this example, we position above the diagonal the 1D colormap encoded PRSPs. Below the diagonal, we

show the 2D color encoded PRSPs with flow lines. This example is only given for demonstration purposes, and SPLOMs are not investigated further within this work.

## 5 Evaluation

To assess our technique, we conducted an online evaluation with 25 volunteers. Prior to the formal evaluation, we conducted the analysis of Section 4, for an initial insight on the focus of the study.

### 5.1 Evaluation Setup

We conducted an anonymous online evaluation, employing the guidelines proposed in the paper of Lam et al. [54]. Initially, we showed a short video to the participants, in which we explain the basic concept behind our approach, without any details concerning the implementation or benefits of the PRSPs. Before the evaluation, we asked demographic questions, regarding the age, vision, previous experience with scatter plots and background of the participants. The main body of the evaluation consisted of five parts, which were controlled user studies measuring user performance (UP) [54] when using our PRSP variant. These five parts were inspired by the tasks, proposed by Sarikaya et al. [2]. In all cases, the participants had to observe static figures and document in written their inputs in a descriptive form. No interaction with the presented figures was required. Figures were shown in adequate quality and size, and in randomized order. Given the limited amount of participants, processing the results was possible with conventional means.

In the first part (*Task 1: Identification of known motif*), we assess whether participants can accurately detect a known motif in the data using PRSPs *(H1)*. We provided users with three datasets containing a known motif or pattern, e.g., a linear trend, or clusters like in Fig. 4, and asked them to identify it in different PRSP variants displayed in randomized order (*"In the following representation, there are three clusters in the data. Are you able to find them?"*). In the second part (*Task 2: Exploration of motifs*), we assess whether participants can accurately find if there is any motif in PRSPs *(H2)*. We provided users with three datasets containing a pattern undisclosed to them, and asked them to identify it in PRSP variants shown in randomized order (*"Do you see a structure in the data? If yes, which?"*). In the third part (*Task 3: Comparison of correlations*), we assess whether participants can accurately compare the level of correlation between PRSPs *(H3)*. We provided users with three pairs from the datasets of Fig. 6, and asked them to identify the one with a higher correlation using three PRSP alternatives in randomized order (*"Which dataset displays a higher level of correlation?"*). The first three tasks did not involve a comparison to conventional scatter or density plots, as we expect from the analysis in Section 4 that scatter plots are already performing well—even better than PRSPs—for these tasks. Still, we need to verify that PRSPs are not hiding such information, due to readability issues. Also, an indication of which PRSP variant is more convenient for the user, is required. Specifically for the third task, we investigate only the comparison of correlations. We expect, given indications from Section 4, that PRSPs are not suitable for reading patterns—instead, only for detecting or comparing motifs.

In the fourth (*Task 4: Cluster detection and density*) and fifth part (*Task 5: Cluster numerosity*), we assess whether participants can more accurately compare the density of clusters in PRSPs than in scatter or density plots *(H4)*, and whether participants can more accurately distinguish the numerosity of clusters in PRSPs than in scatter or density plots *(H5)*, correspondingly. Based on indications from the analysis in Section 4, we expect that our variants may perform better than scatter and density plots for these two tasks, and we would like to confirm this hypothesis. We provided users with, respectively, three and two datasets containing clusters. We asked them to identify the clusters and the denser one, or the cluster with the higher number of data points using PRSP variants, a traditional scatter plot and a density plot in randomized order (*"Do you see any clusters in the data? If yes, which cluster contains a higher number of data points?"* and *"How much more data points does the largest populated cluster contain?"*, respectively).

In summary, the goal of our evaluation is bifold. First of all, we want to show that PRSPs are applicable to the identification, exploration and comparison of certain motifs, where scatter plots and density plots already excel. Second, we aim to show that PRSPs have an advantage in cluster detection, and density and numerosity estimation, over traditional scatter and density plots.

### 5.2 Evaluation Results

We had 25 anonymous participants (gender: 18 men and 7 women, vision: 19 corrected and 6 normal). All, except for five, are related to Visualization and Computer Graphics. They were recruited within our or other collaborating institutes. Most of them have a medium level of experience with scatter plots (11), followed by a high level of experience (8) and, then, novices (6). The participants are between 22 and 45 years of age. They all anonymously volunteered for the evaluation and were not paid for their services. In the remainder of this section, we use the abbreviation *PRSP1* for the PRSP with the 1D colormap encoding, *PRSP2* for the PRSP with the 2D colormap encoding, *PRSP2f* for the PRSP with the 2D colormap encoding and the additional flow lines, *SP* for scatter plots with 0.25 opacity and *DP* for density plots.

*Task 1: Identification of Known Motif.* Using the PRSP1, our evaluation participants were all (25) able to see the predetermined motifs in the data, whether it was the linear trend, the clusters or the manifold shown in Fig. 4. With the PRSP2, they were able to identify the clusters, but not the other motifs. With the PRSP2f, they were able to identify the linear trend and the clusters, but not the manifold. Hence, PRSP1 is suitable for identifying known motifs, while PRSP2 and PRSP2f perform well for clustering, but not for other patterns. Although some noticed an improvement with the flow lines, this was marginal. So, *(H1)* is accepted for PRSP1.

*Task 2: Exploration of Motifs.* Using the PRSP1, our evaluation participants were all (25) able to explore and identify the clusters as shown in Fig. 7 or the pattern in the dataset $(i, j) = (0, 3)$ of the SPLOM in Fig. 8. With the PRSP2, they were able to detect the clusters in the first two cases, but not the pattern. Hence, PRSP1 is suitable for finding motifs, while PRSP2 and PRSP2f perform well for clustering, but for other trends, readability is not possible. Therefore, *(H2)* is accepted for PRSP1.

*Task 3: Comparison of Correlations.* Using the PRSP1, our evaluation participants were all (25) correct in comparing three pairs of data from Fig. 6 and determining which has the highest level of correlation. With the PRSP2, we had between 9 and 13 correct answers out of 25, while with the PRSP2f, we had between 0 and 1 correct answers. Hence, PRSP1 is suitable for comparing correlations and finding the one with a higher degree of correlation, but not the other two. With the introduction of flow lines, the representation becomes more confusing, perhaps due to the readability of colors, or even due to clutter in the screen. One person even commented that PRSP2f are more confusing. So, *(H3)* is accepted for PRSP1.

*Task 4: Cluster Detection and Density*. Using PRSP1, our evaluation participants were all (25) able to detect the clusters and the most dense cluster in the datasets B and D from Fig. 5 and in dataset L from Fig. 7. With the PRSP2, apart from three participants who did not detect even the clusters for dataset D and one who was wrong, 21 detected the clusters and their density correctly. With the SP, for dataset B $23/25$ were correct. For dataset D three did not see the clusters at all, and most $(21/25)$ were wrong in their conclusions. For dataset L most participants were wrong $(22/25)$. With the DP, for dataset B only eight were correct, for dataset D two did not see the clusters at all and the rest were wrong, and for dataset L all participants were wrong. Hence, PRSP1 and PRSP2 are suitable for density detection, while SP and DP present high variations in success rate. So, *(H4)* is accepted for all PRSPs.

*Task 5: Cluster Numerosity*. For dataset B of Fig. 5, with PRSP1 and PRSP2 $16/25$ participants were accurate in their estimation, while another 5 were very close (only 16.7 percent underestimation). With SP, all (but one) participants were inaccurate, with $21/25$ underestimating by 50 percent and 1 overestimating by 50 percent. Two participants even commented that the two clusters are equal in numerosity. With DP, all participants were inaccurate, with $5/25$ underestimating by 50 percent, $19/25$ thinking that the clusters had same numerosity, and one person not being able to answer. For dataset L of Fig. 7, with PRSP1 and PRSP2 participants were more accurate in their estimation than SP and DP, with most people slightly overestimating their prediction. SP and DP had larger variations in the responses, with the participants overestimating the numerosity—up to saying that the larger cluster is 20 times larger. Also, 3 people for SP and 4 for DP could not give a conclusive answer. Hence, SP and DP are inaccurate or inconclusive regarding numerosity tasks, but with PRSP1 and PRSP2 we obtain a higher accuracy and precision in responses than SP and DP. Thus, *(H5)* is accepted for all PRSP variants.

## 6 Discussion

Scatter plots excel at certain tasks related to data spatialization, such as detecting patterns or clusters. Although in PRSPs, certain data motifs—in particular clusters—can be identified, the representation was not explicitly designed for this purpose or with the intention of replacing traditional scatter plots. Instead, PRSPs should be seen as an auxiliary data visualization technique, which acts complementarily to traditional scatter or density plots, avoiding overplotting through pixel-based, space-filling mechanisms. As illustrated in Sections 4 and 5, PRSPs and, especially, PRSPs with the 1D colormap, are primarily suitable for comparing correlations, for cluster density identification and numerosity detection.

We confirmed in our evaluation, in Section 5, that PRSPs provide a more accurate representation of the density of the datasets than the traditional scatter plots—even with opacity—and their respective density plots. In the design of the relaxation encoding, we chose color as a visual variable with strong selective and associative properties [55]. This selection enables the identification and localization of patterns, their densities and numerosities. It helps to convey information about the size of structures and the patterns. Initial analysis showed us that inverting the 1D colormap to display the distortion, as shown in Fig. 2, helps to better display motifs and patterns. The user's attention is more drawn towards the brighter data items denoted with yellow, which are the less displaced data items. This makes the technique particularly suitable for data with strong distortions.

In Section 4.3, we arranged our PRSPs in a SPLOM configuration to demonstrate applicability to more than two dimensions in a proof-of-concept. In this case, our approach could be used as an initial step for an extension towards scagnostics [56], or for the design of novel, aggregative, abstracted views on the data, which can be used to save screen space. This should take into account approaches that focus both on automatic extraction of local and global motifs within plots, such as the recent work of Matute [57] and Shao [58]. It would also be interesting to take advantage of algorithms from the field of image processing, for better motif or pattern detection—possibly, through edge detection.

Scalability to larger datasets has been demonstrated in Section 4.2. Cases, where the number of data items is larger than the available screen pixels, requires downsampling and interaction techniques to display the data and to provide exploratory means, such as zooming and panning. Standard image downsampling techniques that create an image pyramid, which is displayed at the required level, should perform well in most cases. However, it can be more beneficial to adjust the downsampling to the intended goal of the analyst. For example, keeping at each downsampling step the item that differs the most from the rest for the next coarser level could be used to highlight outliers. If the goal is to preserve motifs or cluster boundaries, then the vector median is a good choice. It is important that the downsampling is performed before adding glyphs, such as the flow lines, in order to avoid clutter. In these cases, the flow lines will be computed once for each image region of size $a \times a$, where $a$ can be a user-defined value. As flow lines were considered in general confusing in the evaluation, a useful improvement would be to use bundling [59] to reduce clutter. Other techniques, such as Fisheye lenses [36], could also be employed.

We demonstrated in Section 4.1 that our PRSPs remove clutter due to overplotting by performing an optimal 1:1 mapping of the point-to-pixel positions through a linear sum assignment. Due to its long computation times, we also tested a median-split mapping, as a fast alternative to the linear sum assignment. However, the recursive subdivision at the median positions can become noticeable and corrupt the motifs in the data, to some degree. Simple relaxation, such as collision detection and repelling algorithms, as used in cartogram-based distortion [60] are not suitable in our approach, as the target domain is not unbounded and should be completely filled. This would lead to long computation times, in our case. However, approaches like the ones proposed by

Gale and Shapley [61], or Gusfield and Irving [62] would be interesting alternatives to the linear sum assignment. Also, fast solvers of the linear assignment problem are just becoming more widely available and could be used to replace the median-split mapping, in the future.

Our evaluation focused on a comparison between the proposed PRSP variants, and to some extent against scatter and density plots. Yet, there are a lot of other techniques solving different aspects within the overplotting topic, as mentioned in Section 2. In particular, in the latter work of Micallef et al. [4], the authors propose to reduce overplotting in scatter plots through the construction of a cost function. This function captures visual aspects, such as marker size and opacity, aspect ratio, color, and rendering order, to optimize the design of a scatter plot for a particular task. However, as also remarked by the authors, balancing terms in the cost function needs to be done carefully with respect to different tasks or data (e.g., large and opaque markers are beneficial for outlier detection and class separation, respectively). On the positive side, this makes their approach extensible to different kind of tasks, while PRSPs have a more narrow task target. On the negative side, the optimization can be challenging, while PRSPs do not require any particular fine-tuning, apart from the mapping and encoding selection, being more simple. Additionally, optimizing the visualization parameters [4] mitigates overplotting and makes the visual design of scatter plots more readable, but the inherent problem of overplotting cannot be solved entirely by such approaches. Regarding correlation detection, we only evaluated relative assessments for our PRSP. Absolute correlation assessments are often of higher interest for practical use. This should be further investigated following the setup of Li et al. [63]. Finally, a more thorough comparison with traditional scatter and density plots should be conducted, regarding suitability for other data analysis tasks. This is out of scope for our present work, but a thorough study with additional techniques—in particular splatterplots [24], or perceptually optimized scatter plots [4]—should be conducted, in the future.

Another potential shortcoming of the current evaluation is that it was conducted with a limited amount of participants, who were mostly knowledgeable about scatter plots. Basic knowledge in visualization was desired, so that we could ensure familiarity with basic representations such as scatter plots and density plots—as the goal was to investigate if PRSPs provide benefits for certain tasks. Therefore, the participants were all working, or used to work, in a visualization or computer graphics environment, or were students that already attended a visualization course. A crowdsourcing evaluation approach [64] should be preferred in the future, in order to include also a more general public, i.e., additional novice users, and to render the results of the evaluation more significant.

The analysis of our results in Section 4 and the evaluation in Section 5 have also exposed some limitations of our approach. To begin with, spatialization of data points is the most effective visual variable in scatter plots, according to Bertin [65]. In our case, we are changing the spatial location making it potentially less useful. However, by minimizing the displacement of pixels with respect to their original position and carefully replacing the lost spatial information with color, we were able to preserve discriminative motifs within the data—in particular, clusters and especially with the use of the PRSPs with the 1D color encoding. We recognize that the PRSPs on their own are not as intuitive as scatter plots and require some familiarization. Deriving the absolute position of data points is fundamentally hard and error-prone to some degree; hence, we propose to use PRSPs only as a supplement and not as a replacement to traditional scatter plots. Nevertheless, if no traditional scatter plot is available, experienced or familiarized users might still be able to solve positional tasks. The 1D colormap preserves motifs and cluster centers well, while smooth color changes indicate similar origins.

Although it seldomly happened, the linear sum assignment may create single outliers due to the euclidean distance optimization. To solve this, incorporating additional parameters in the objective function might be useful, such as minimizing the maximal displacement instead of the euclidean distance. However, we have not evaluated this so far, and propose it as future work. Also, the mapping from the scatter plot space to the PRSP space is not bijective. Theoretically, the same PRSP could exist for different data sets. However, this is only a theoretical problem and would occur only in synthetic cases. Additionally, a minor issue is that some pixels in the representation may be unused if $N \neq n \times m$. This might entail the limitation of losing screen space.

Additional point-to-pixel mapping and relaxation encoding solutions were regarded, which we propose here as points for future work. Quad-tree-based density ordering algorithms or $k$-space filling algorithms that take data point density into account can be surrogates, for the currently employed point-to-pixel mappings. In the relaxation mapping, calculating the distance of each pixel to the $n$th nearest neighbor in the dataset might provide better insight into clustered data structures. Using line integral convolution (LIC) representations might also give a good notion of the relaxation introduced in the resulting PRSPs [66]. Finally, the use of animation to help the transition between the two spaces would provide the user with a seamless, dynamic and efficient way of exploring the data in both spaces, simultaneously. Other interactivity could also prove helpful for pattern identification, structure density and cluster size information retrieval—including new brushing approaches.

## 7 Conclusion

Scatter plots are simple, but powerful, visual representations of two-dimensional data with the ability to communicate the existence of data clusters, correlations, or outliers. However, with an increasing number of data points, they become less effective, due to commonly occurring overplotting. In this paper, we presented the Pixel-Relaxed Scatter Plots (PRSPs), a pixel-based, space-filling variant. Data points of the dataset are mapped uniquely to image pixels, using different mapping solutions to avoid the overlap of data points and different ways of calculating and encoding the introduced relaxation. We demonstrated the results of our approach with several synthetic and realistic datasets, and we discussed the applicability and suitability of our variant to different cases and tasks. We conducted a user evaluation with 25 participants, to confirm the advantages and to identify the limitations of our approach. The PRSP representation can be useful for searching for particular known motifs in the data, for

exploring the depicted datasets, for comparing the density and numerosity in different regions of the representation and, possibly, for characterizing distributions and determining their level of correlation.

## Acknowledgments

This paper was partly written in collaboration with the VRVis Competence Center. VRVis is funded by BMVIT, BMWFW, Styria,SFG and Vienna Business Agency in the scope of COMET - Competence Centers for Excellent Technologies (854174), managed by FFG. The authors thank A. Amirkhanov, M. Waldner and H.-Y. Wu from TU Wien for their input, and the anonymous evaluation participants.

## References

[1] M. Friendly and D. Denis, "The early origins and development of the scatterplot," *J. Hist. Behavioral Sci.*, vol. 41, no. 2, pp. 103–130, 2005.
[2] A. Sarikaya and M. Gleicher, "Scatterplots: Tasks, data, and designs," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 402–412, Jan. 2018.
[3] M. Behrisch, M. Blumenschein, N. W. Kim, L. Shao, M. El-Assady , J. Fuchs, D. Seebacher, A. Diehl, U. Brandes, H. Pfister, et al., "Quality metrics for information visualization," in *Computer Graphics Forum*, vol. 37, Hoboken, NJ, USA: Wiley, 2018, pp. 625–662.
[4] L. Micallef, G. Palmas, A. Oulasvirta, and T. Weinkauf, "Towards perceptual optimization of the visual design of scatterplots," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 6, pp. 1588–1599, Jun. 2017.
[5] G. Ellis and A. Dix, "A taxonomy of clutter reduction for information visualisation," *IEEE Trans. Vis. Comput. Graph.*, vol. 13, no. 6, pp. 1216–1223, Nov.-Dec. 2007.
[6] A. Dix and G. Ellis, "By chance enhancing interaction with large data sets through statistical sampling," in *Proc. Work. Conf. Adv. Visual Interfaces*, 2002, pp. 167–176.
[7] G. Ellis, E. Bertini, and A. Dix, "The sampling lens: Making sense of saturated visualisations," in *Proc. Extended Abstracts Human Factors Comput. Syst.*, 2005, pp. 1351–1354.
[8] E. Bertini and G. Santucci, "Give chance a chance: Modeling density to enhance scatter plot quality through random data sampling," *In. Vis.*, vol. 5, no. 2, pp. 95–110, 2006.
[9] H. Chen, W. Chen, H. Mei, Z. Liu, K. Zhou, W. Chen, W. Gu, and K.-L. Ma, "Visual abstraction and exploration of multi-class scatterplots," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1683–1692, Dec. 2014.
[10] M. C. Stone, K. Fishkin, and E. A. Bier, "The movable filter as a user interface tool," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 1994, pp. 306–312.
[11] D. Brodbeck, M. Chalmers, A. Lunzer, and P. Cotture, "Domesticating bead: adapting an information visualization system to a financial institution," in *Proc. Vis. Conf. Inf. Vis. Symp. Parallel Rendering Symp.*, 1997, pp. 73–80.
[12] C. Ahlberg and B. Shneiderman, "Visual information seeking: Tight coupling of dynamic query filters with starfield displays," in *The Craft of Information Visualization*. Amsterdam, The Netherlands: Elsevier, 2003, pp. 7–13.
[13] K. Chen and L. Liu, "A visual framework invites human into the clustering process," in *Proc. 15th Int. Conf. Sci. Statistical Database Manag.*, 2003, pp. 97–106.
[14] J. Li, J. J. van Wijk, and J.-B. Martens, "Evaluation of symbol contrast in scatterplots," in *Proc. IEEE Pacific Vis. Symp.*, 2009, pp. 97–104.
[15] C. Waldeck and D. Balfanz, "Mobile liquid 2D scatter space (ML2DSS)," in *Proc. 8th Int. Conf. Inf. Vis.*, 2004, pp. 494–498.
[16] A. Woodruff, J. Landay, and M. Stonebraker, "Constant density visualizations of non-uniform distributions of data," in *Proc. 11th Annu. ACM Symp. User Interface Softw. Technol.*, 1998, pp. 19–28.
[17] J. Li, J.-B. Martens, and J. J. van Wijk, "A model of symbol size discrimination in scatterplots," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2010, pp. 2553–2562.
[18] J. Matejka, F. Anderson, and G. Fitzmaurice, "Dynamic opacity optimization for scatter plots," in *Proc. 33rd Annu. ACM Conf. Human Factors Comput. Syst.*, 2015, pp. 2707–2710.
[19] J. Li, J. J. van Wijk, and J.-B. Martens, "A model of symbol lightness discrimination in sparse scatterplots," in *Proc. IEEE Pacific Vis. Symp.*, 2010, pp. 105–112.
[20] R. Kosara, S. Miksch, and H. Hauser, "Focus+context taken literally," *IEEE Comput. Graph. Appl.*, vol. 22, no. 1, pp. 22–29, Jan./Feb. 2002.
[21] J. Staib, S. Grottel, and S. Gumhold, "Enhancing scatterplots with multi-dimensional focal blur," *Comput. Graph. Forum*, vol. 35, no. 3, pp. 11–20, 2016.
[22] D. B. Carr, R. J. Littlefield, W. Nicholson, and J. Littlefield, "Scatterplot matrix techniques for large N," *J. Amer. Statistical Assoc.*, vol. 82, no. 398, pp. 424–436, 1987.
[23] W. S. Cleveland, M. E. McGill, and R. McGill, "The shape parameter of a two-variable graph," *J. Amer. Statistical Assoc.*, vol. 83, no. 402, pp. 289–300, 1988.
[24] A. Mayorga and M. Gleicher, "Splatterplots: Overcoming overdraw in scatter plots," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 9, pp. 1526–1538, Sep. 2013.
[25] B. Li, D. A. Griffith, and B. Becker, "Spatially simplified scatterplots for large raster datasets," *Geo-Spatial Inf. Sci.*, vol. 19, no. 2, pp. 81–93, 2016.
[26] C. Collins, G. Penn, and S. Carpendale, "Bubble sets: Revealing set relations with isocontours over existing visualizations," *IEEE Trans. Vis. Comput. Graph.*, vol. 15, no. 6, pp. 1009–1016, Nov.-Dec. 2009.
[27] M. Tory, D. Sprague, F. Wu, W. Y. So, and T. Munzner, "Spatialization design: Comparing points and landscapes," *IEEE Trans. Vis. Comput. Graph.*, vol. 13, no. 6, pp. 1262–1269, Nov.-Dec. 2007.
[28] M. C. Hao, U. Dayal, R. K. Sharma, D. A. Keim, and H. Janetzko, "Variable binned scatter plots," *Inf. Vis.*, vol. 9, no. 3, pp. 194–203, 2010.
[29] T. Ledl, "Kernel density estimation: Theory and application in discriminant analysis," *Austrian J. Statistics*, vol. 33, no. 3, pp. 267–279, 2004.
[30] D. A. Keim and A. Herrmann, "The gridfit algorithm: An efficient and effective approach to visualizing large amounts of spatial data," in *Proc. Vis.*, 1998, pp. 181–188.
[31] M. Trutschl, G. Grinstein, and U. Cvek, "Intelligently resolving point occlusion," in *Proc. IEEE Symp. Inf. Vis.*, 2003, pp. 131–136.
[32] T. N. Dang, L. Wilkinson, and A. Anand, "Stacking graphic elements to avoid over-plotting," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1044–1052, Nov.-Dec. 2010.
[33] Y. K. Leung and M. D. Apperley, "A review and taxonomy of distortion-oriented presentation techniques," *ACM Trans. Comput.-Human Interaction*, vol. 1, no. 2, pp. 126–160, 1994.
[34] M. Sarkar, S. S. Snibbe, O. J. Tversky, and S. P. Reiss, "Stretching the rubber sheet: a metaphor for viewing large layouts on small screens," in *Proc. 6th Annu. ACM Symp. User Interface Softw. Technol.*, 1993, pp. 81–91.
[35] M. S. T. Carpendale, D. J. Cowperthwaite, and F. D. Fracchia, "3-dimensional pliable surfaces: For the effective presentation of visual information," in *Proc. 8th Annu. ACM Symp. User Interface Softw. Technol.*, 1995, pp. 217–226.
[36] M. Sarkar and M. H. Brown, "Graphical fisheye views," *Commun. ACM*, vol. 37, no. 12, pp. 73–83, 1994.
[37] J. Stasko, R. Catrambone, M. Guzdial, and K. McDonald, "An evaluation of space-filling information visualizations for depicting hierarchical structures," *Int. J. Human-Comput. Studies*, vol. 53, no. 5, pp. 663–694, 2000.
[38] B. Johnson and B. Shneiderman, "Tree-maps: A space-filling approach to the visualization of hierarchical information structures," in *Proc. 2nd Conf. Vis.*, 1991, pp. 284–291.
[39] D. A. Keim, M. C. Hao, U. Dayal, H. Janetzko, and P. Bak, "Generalized scatter plots," *Inf. Vis.*, vol. 9, no. 4, pp. 301–311, 2010.
[40] S. Bachthaler and D. Weiskopf, "Continuous scatterplots," *IEEE Trans. Vis. Comput. Graph.*, vol. 14, no. 6, pp. 1428–1435, Nov.-Dec. 2008.
[41] D. A. Keim, "Designing pixel-oriented visualization techniques: Theory and applications," *IEEE Trans. Vis. Comput. Graph.*, vol. 6, no. 1, pp. 59–78, Jan.-Mar. 2000.
[42] D. A. Keim, J. Schneidewind, and M. Sips, "Scalable pixel based visual data exploration," in *Pixelization Paradigm*. Berlin, Germany: Springer, 2007, pp. 12–24.
[43] D. A. Keim, M. Ankerst, and H.-P. Kriegel, "Recursive pattern: A technique for visualizing very large amounts of data," in *Proc. 6th Conf. Vis.*, 1995, Art. no. 279.
[44] D. A. Keim, M. C. Hao, U. Dayal, and M. Hsu, "Pixel bar charts: A visualization technique for very large multi-attribute data sets," *Inf. Vis.*, vol. 1, no. 1, pp. 20–34, 2002.

[45] J.-D. Fekete and C. Plaisant, "Interactive information visualization of a million items," in *The Craft of Information Visualization*. Berlin, Germany: Elsevier, 2003, pp. 279–286.
[46] H. W. Kuhn, "The Hungarian Method for the Assignment Problem," *Naval Res. Logistics Quart.*, vol. 2, no. 1–2, pp. 83–97, Mar. 1955.
[47] J. Bernard, M. Steiger, S. Mittelstädt, S. Thum, D. Keim, and J. Kohlhammer, "A survey and task-based quality assessment of static 2D colormaps," in *Proc. Vis. Data Anal.*, 2015.
[48] S. Bremm, T. v. Landesberger, J. Bernard, and T. Schreck, "Assisted descriptor selection based on visual comparative data analysis," *Comput. Graph. Forum*, vol. 30, no. 3, pp. 891–900, 2011.
[49] R. Borgo, J. Kehrer, D. H. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen, "Glyph-based visualization: Foundations, design guidelines, techniques and applications," in *Proc. EuroGraphics Conf.*, 2013, pp. 39–63.
[50] XmdvTool Datasets, [Online]. Available: http://davis.wpi.edu/xmdv/datasets.html, Accessed on: 01.09.2018
[51] S. Bremm, M. He, T. V. Landesberger, and D. W. Fellner, "PCDC - on the highway to data - a tool for the fast generation of large synthetic data sets," in *Proc. Int. Workshop Visual Analytics*, 2012.
[52] P. A. Tukey and J. W. Tukey, "Graphical display of data sets in three or more dimensions," in *Interpreting Multivariate Data*, Hoboken, NJ, USA: Wiley, 1981, pp. 189–275.
[53] J.-F. Im, M. J. McGuffin, and R. Leung, "GPLOM: The generalized plot matrix for visualizing multidimensional multivariate data," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, pp. 2606–2614, Dec. 2013.
[54] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale, "Empirical studies in information visualization: Seven scenarios," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 9, pp. 1520–1536, Sep. 2012.
[55] M. Carpendale, "Considering visual variables as a basis for information visualisation," Dept. Comput. Sci., Univ. Calgary, Calgary, Canada, Rep. TR 2001–693-16, 2003.
[56] L. Wilkinson, A. Anand, and R. Grossman, "Graph-theoretic scagnostics," in *Proc. IEEE Symp. Inf. Vis.*, 2005, pp. 157–164.
[57] J. Matute, A. C. Telea, and L. Linsen, "Skeleton-based scagnostics," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 542–552, Jan. 2018.
[58] L. Shao, T. Schleicher, M. Behrisch, T. Schreck, I. Sipiran, and D. A. Keim, "Guiding the exploration of scatter plot data using motif-based interest measures," *J. Visual Languages Comput.*, vol. 36, pp. 1–12, 2016.
[59] D. Holten and J. J. Van Wijk, "Force-directed edge bundling for graph visualization," in *Computer Graphics Forum*, vol. 28, no. 3. Hoboken, NJ, USA: Wiley, 2009, pp. 983–990.
[60] C. Panse, M. Sips, D. Keim, and S. North, "Visualization of geospatial point sets via global shape transformation and local pixel placement," *IEEE Trans. Vis. Comput. Graph.*, vol. 12, no. 5, pp. 749–756, Sep.-Oct. 2006.
[61] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *Amer. Math. Monthly*, vol. 69, no. 1, pp. 9–15, 1962.
[62] D. Gusfield and R. W. Irving, *The Stable Marriage Problem: Structure and Algorithms*. Cambridge, MA, USA: MIT Press, 1989.
[63] J. Li, J.-B. Martens, and J. J. Van Wijk, "Judging correlation from scatterplots and parallel coordinate plots," *Inf. Vis.*, vol. 9, no. 1, pp. 13–30, 2010.
[64] R. Borgo, L. Micallef, B. Bach, F. McGee, and B. Lee, "Information visualization evaluation using crowdsourcing," in *Computer Graphics Forum*, vol. 37, Hoboken, NJ, USA: Wiley, 2018, pp. 573–595.
[65] J. Bertin, "Sémiologie graphique: Les diagrammes – les réseaux – les cartes [Semiology of graphics]," 1967.
[66] B. Cabral and L. C. Leedom, "Imaging vector fields using line integral convolution," in *Proc. 20th Annu. Conf. Comput. Graphics Interactive Techn.* 1993, pp. 263–270.

**Renata Georgia Raidou** received the PhD degree in medical visualization from the Eindhoven University of Technology, the Netherlands—for which she received the Dirk Bartz Prize for Visual Computing in Medicine (1st Place) at Eurographics 2017, and the EuroVis Best PhD Award 2018. She is a postdoctoral researcher at TU Wien, Austria. Her research interests include the interface between Visual Analytics, Image Processing and Machine Learning, with a strong focus on medical applications.

**M. Eduard Gröller** is professor at TU Wien, Austria, and adjunct professor of computer science with the University of Bergen, Norway. His research interests include computer graphics, visualization and visual computing. He became a fellow of the Eurographics association in 2009. He is the recipient of the Eurographics 2015 Outstanding Technical Contributions Award.

**Martin Eisemann** is a full professor with the TH Köln since 2015. Between 2007 and 2014 he was a (visiting) researcher at several institutions including TU Braunschweig, TU Delft, Saarland University, EDM Hasselt and the Max-Planck Institute für Informatik. His main research interests include image- and video-based rendering and editing, visual analytics, and realistic and interactive rendering.