Reducing Ambiguities in Line-based Density Plots by Image-space Colorization

Yumeng Xue, Patrick Paetzold, Rebecca Kehlbeck, Bin Chen, Kin Chung Kwan, Yunhai Wang, and Oliver Deussen

Abstract—Line-based density plots are used to reduce visual clutter in line charts with a multitude of individual lines. However, these traditional density plots are often perceived ambiguously, which obstructs the user's identification of underlying trends in complex datasets. Thus, we propose a novel image space coloring method for line-based density plots that enhances their interpretability. Our method employs color not only to visually communicate data density but also to highlight similar regions in the plot, allowing users to identify and distinguish trends easily. We achieve this by performing hierarchical clustering based on the lines passing through each region and mapping the identified clusters to the hue circle using circular MDS. Additionally, we propose a heuristic approach to assign each line to the most probable cluster, enabling users to analyze density and individual lines. We motivate our method by conducting a small-scale user study, demonstrating the effectiveness of our method using synthetic and real-world datasets, and providing an interactive online tool for generating colored line-based density plots.

Index Terms—Trajectory data, time series, density-based visualization, clustering, coloring

1 INTRODUCTION

Line-based plots are a popular way to visualize time series data [20] and trajectory data [10]. These plots are widely used across domains, such as finance, healthcare, and navigation, to highlight evolving trends in complex datasets. To emphasize patterns and gain insights, two effective techniques are line coloring and density plots. By using line coloring, different lines in a plot can be assigned distinct colors, making it easier to distinguish between different groups or categories. This technique is particularly useful when dealing with multiple time series or trajectories that are plotted together. Alternatively, density plots show the distribution of data at each point in time or space. These plots use color or shading to indicate areas of high or low data density, providing a more detailed picture of the underlying patterns in the data.

Line coloring assigns a color to each line. However, even in small datasets, line intersections can cause overplotting, where lines are drawn on top of each other, making it difficult to interpret the data. This problem becomes more pronounced in complex datasets with a larger number of lines, resulting in visual clutter that can obscure underlying trends [30], see Fig. 1a. Mainly, this has been addressed by replacing individual lines through visual abstractions [17, 42, 73], which use abstract glyphs to display statistical information about the lines.

Using density plots [34, 44] shifts the focus from visualizing individual data items to aggregating data attributes in image space, communicating the structure of the data better. By using a continuous color map to visually communicate the density at a given position, density plots can effectively reduce visual clutter. However, the individual line information is lost. Such plots do not reveal which and how lines interact at a specific position. To alleviate this issue, interaction techniques have to be used (e.g., timebox [23]) with representative lines (such as kd-box [81]) being used to explore patterns of interest.

In general, due to the continuous nature of time series and trajectory data, line-based density plots are more challenging to interpret than, e.g., point-based density plots. Visual ambiguity within the perceived patterns often leads to false conclusions, as the Gestalt principles of good continuation and similarity strongly influence the perception of interacting lines. Similarly, the Principle of Unambiguous Data Depiction introduced by Kindlmann and Scheidegger [33] describes visualiza-
We demonstrate the capabilities of our method using selected synthetic data that lead humans to draw wrong conclusions. Feng et al. [34] extended KDE to curves and introduced clustering to construct curve-boxplot-like abstractions of multiple line data and enable various analyses, including clustering, abstraction, and interactive exploration. Clustering of lines can be achieved by applying these similarity metrics. In addition to the previously introduced clustering approaches that either visually cluster lines by color or abstract them, edge bundling techniques [24, 25, 66] alter the course of individual lines to combine them into bundles to reduce visual clutter. To incorporate geographical restrictions like roads, specially tailored edge bundling techniques [62, 79] were proposed. Recently Wallinger et al. [68, 69] proposed an edge-path bundling method that reduced the ambiguities introduced by edge bundling. Our method is focused on reducing the ambiguity while not altering the original data, as edge bundling does.

Interaction methods are also important for analyzing lines. For example, QuerySketch [71] allows users to sketch freely and query the data lines that match the shape of the sketched line. Hurter et al. [27] introduced interactive paradigms to extract trajectories of interest from large-scale airplane trajectory data. Hochheiser and Shneiderman [23] presented Timebox to query data that passes through a box, representing a range of positions within a certain time. Recently, Zhao et al. [81] proposed KD-Box by using a KD-tree to speed up the timebox query and introduce representative lines to assist analysts in exploring the local pattern details of interest. However, these methods require significant user interaction and learning costs, and the small number of representative lines may not fully reflect the overall composition of patterns.

2.3 Line-based Analysis

Line-based data analysis methods use lines as fundamental units of data and enable various analyses, including clustering, abstraction, and interactive exploration. Line clustering uses different metrics [80] to measure the distance between 2D or 3D lines, such as the Euclidean distance [11, 55], curvature or torsion [41, 75], and user-specified streamline predicates [56]. Dynamic time warping (DTW) [45] is another widely used metric to measure similarity between time series data. The clustering of lines can be achieved by applying these similarity metrics. To obtain cluster centers for clustering time series data, Petitjean et al. [49] proposed a global averaging method based on DTW, while Gaffney and Smyth [21] presented a framework for clustering curves based on probabilistic curve alignment models. Some methods use Euclidean space vectors to represent the line curves and then apply Euclidean distance-based clustering methods [9, 55].

However, even after obtaining clustering information, further analysis remains a challenge because lines can still be cluttered, and clustering methods are sensitive to noise. Uncertainty visualizations therefore use glyphs to visualize statistical information and enable overviews. Mirzargar et al. [42] generalized the curve boxplot for ensembles of curves by introducing a functional band depth. Ferstl et al. [17] applied clustering to construct curve-boxplot-like abstractions of multiple line bundles. Palmas et al. [48] abstracted clusters in parallel coordinates through edge bundling, facilitating further interaction with clusters. These visualization methods are effective only for regularly distributed lines and are not suitable for noisy datasets.

In addition to the previously introduced clustering approaches that either visually cluster lines by color or abstract them, edge bundling techniques [24, 25, 66] alter the course of individual lines to combine them into bundles to reduce visual clutter. To incorporate geographical restrictions like roads, specially tailored edge bundling techniques [62, 79] were proposed. Recently Wallinger et al. [68, 69] proposed an edge-path bundling method that reduced the ambiguities introduced by edge bundling. Our method is focused on reducing the ambiguity while not altering the original data, as edge bundling does.

Interaction methods are also important for analyzing lines. For example, QuerySketch [71] allows users to sketch freely and query the data lines that match the shape of the sketched line. Hurter et al. [27] introduced interactive paradigms to extract trajectories of interest from large-scale airplane trajectory data. Hochheiser and Shneiderman [23] presented Timebox to query data that passes through a box, representing a range of positions within a certain time. Recently, Zhao et al. [81] proposed KD-Box by using a KD-tree to speed up the timebox query and introduce representative lines to assist analysts in exploring the local pattern details of interest. However, these methods require significant user interaction and learning costs, and the small number of representative lines may not fully reflect the overall composition of patterns.

2.4 Color Mapping

Color maps are an important aspect of data visualization, as they have a significant impact on how effectively the viewer perceives and understands the underlying data. Zhou and Hansen [82] conducted a
We restrict our consideration of trajectories to one and two-dimensional representations of more abstract spaces, like eye-tracking trajectories. For example, geospatial data, like traffic data or human movements, or time series representation. Two-dimensional co-domains can represent, for instance, the density distribution of a surface or a map. For one-dimensional co-domains, the trajectory is a time series representation. Two-dimensional co-domains can represent, for example, geospatial data, like traffic data or human movements, or more abstract spaces, like eye-tracking trajectories.

3.2 Ambiguities in Density Plots

While phenomena like visual ambiguities and visual artifacts introduced by scatter plots [40] and line-based plots have been studied [51], little research has been done on density plots. Compared to density representations of scatter plots, which are straightforward to interpret, the interplay of multiple lines in a line-based density plot might add visual ambiguities hampering their clarity. Investigating line-based density plots of real-world datasets, we found patterns with ambiguities and further compiled a list of three abstracted types of patterns (Fig. 2).

We focus ourselves on these three recurring patterns in our investigation and do not claim the list to be complete. In Section 6, we show the results of our user study, underpinning that users perceive line-based density plots ambiguously. The upper rows of each subfigure show possibilities of individual trends that, when combined, form a similar pattern. Even if only two trends are superimposed, it is challenging to decide which individual trends the density plot is composed of.

Illusory patterns. The combined density plot in the lower row of Fig. 2a is visually dominated by the central horizontal line. An observer could assume this dominant region is superimposed on a less dense but noisy region surrounding it. This perceived continuation is also supported by the Gestalt principle of good continuity [72]. Our example illustrates that this perception is deceptive, as the observed pattern is actually a composition of the two patterns on the top left. This ambiguity is often present in real-world datasets, such as ship trajectories before and after passing through a strait (see Fig. 6). We refer to patterns that appear to have a high-density trend only in the combined line-based density plot that is not present in the individual underlying patterns as “illusory patterns”.

Ambiguous continuation. If multiple independent, well-separated trends partly overlap with a similar orientation in the overlapping region, their continuation is ambiguous. In the lower row of Fig. 2b, it is not clear how the individual trends continue after intersecting. Both combinations on the top could create such a pattern. Thus, their continuation remains unclear. This ambiguity frequently appears in data we have observed, such as time series data with crossovers and overlaps between trends (see Fig. 4).

Disconnected clusters. Dense regions in plots attract attention. If a line-based density plot contains multiple dense regions, it is ambiguous if they belong to the same trend. A line-based density plot with two such dense regions is shown in the lower row of Fig. 2c. Due to their visual saliency and similar density, they appear to be related. Perceptual psychology explains this with the Gestalt principle of similarity. However, this impression is deceptive since the pattern is actually composed of the two completely independent patterns on the top of the upper row of Fig. 2c. Thus it is ambiguous if the combined density plot consists of two line bundles fanning out in the center of the plot or two independent trends. It is common to see disconnected clusters in real-world datasets; an example of this can be observed in Fig. 5. Even these straightforward examples show that interpreting line-based
density plots is challenging, and identifying the underlying trends is often ambiguous. Patterns in real-world data are even more complex than our examples. We acknowledge the possibility of the existence of other types of ambiguities, which is an interesting topic for further research. Already just the three types which we identified show that there is a need to visually better convey trends and patterns. This is the motivation for us to propose our density plot coloring method that helps to identify individual patterns within line-based density plots. Our method assigns similar colors to regions with similar lines passing through (Figs. 8a, 8c, and 8e), which helps users intuitively understand the true composition of the related line patterns and their relationships.

4 Method

Our goal is to reduce ambiguities in line-based density plots, identifying and highlighting similar regions. Previously introduced data-space clustering methods cluster individual data items, like lines or scatter points. In contrast, we apply clustering in the image-space, grouping regions based on the lines passing through them.

Generally, creating density plots involves three main steps: discretizing the canvas into 2D bins, counting the data items that touch each bin, calculating the bin’s density, and optionally applying a smoothing method, e.g., a kernel method. This bin-summarize-smooth paradigm was introduced by Wickham [74]. Our method assigns colors to bins in the image space based on their similarity. Similar bins are grouped into clusters with hue values assigned to them.

Our pipeline (see Fig. 3) consists of four main steps: bin-based feature set extraction, reduction of the number of bins by sampling and thresholding, hierarchical clustering, and finally, the assignment of colors to bins. We implemented our method as a web-based analysis tool that offers various user-controllable parameters to analyze the data and steer the colorization.

4.1 Feature Set Extraction

As a first step, we discretize the plot’s canvas into bins. We see a bin as an atomic unit of the density plot, which can contain multiple pixels, but each bin can only have one color, i.e., the number of bins can be considered as the logical resolution. For simplicity, in the following, a bin is equivalent to a pixel. We define a feature set for each bin that contains the identifiers of all lines touching it. This feature set expresses the relationship between a bin p and every line \( L = \{ L_1, L_2, \ldots, L_N \} \) in the dataset. It encodes which lines of \( L \) are spatially close to the center of \( p \). For each bin, we obtain a feature set \( S(p) \):

\[
S(p) = \{ i \mid L_i \in L \land \text{distance}(L_i, p) < T \}.
\]

where distance\((L_i, p)\) is the closest distance between line \( L_i \) and center of bin \( p \). The parameter \( T \) is conceptually the radius of a disc positioned at the center of \( p \), defining the extraction area for that bin. As default, we set \( T = 1 \). Due to the circular distance measure, the disc created by \( T \) may cover a slightly larger area than the bin. The feature set \( S(p) \) then contains the identifiers of all lines running through the disc at position \( p \). Increasing \( T \) expands the extraction area of bins, which will result in more similar feature sets for adjacent bins.

4.2 Bin Sampling

If the bins correspond to individual pixels, the number of feature sets \( S(p) \) is equal to the number of pixels in the density plot. Even a medium-resolution density plot with a resolution of 1000 × 500 pixels contains 500,000 bins. In the subsequent step of our pipeline, we use agglomerative hierarchical clustering [29] to group bins based on their similarity. It has a time complexity of \( O(n^2 \log(n)) \), which makes it challenging to use it to cluster such a large number of bins. One potential way to reduce the number of bins is to increase the bin size. However, this inevitably sacrifices details and results in a loss of resolution. Alternatively, we can either sample bins or exclude bins of low-density regions. Using sampling has the advantage that we can exploit the continuous nature of the line-based datasets that underlay the visualization - feature sets of neighboring bins are typically similar due to the spatial continuity of lines. Thus sampling does roughly retain the distribution of the feature sets. In addition, users are typically interested in sufficiently dense regions. Low-density bins (e.g., those touching fewer than ten lines) typically contain a lot of noise and are less noticeable to the user. To filter out such bins, we provide a user-customizable minimum density threshold.

4.3 Hierarchical Clustering

We group the sampled bins using agglomerative hierarchical clustering. To cluster the bins, we have to use a distance metric that is appropriate for set-type data. Commonly employed are the Jaccard index \( I(A, B) = \frac{|A \cap B|}{|A| + |B|} \) and Sørensen-Dice coefficient \( DSC(A, B) = \frac{2|A \cap B|}{|A| + |B|} \) for sets \( A \) and \( B \). A ratio close to 1 indicates a high degree of similarity. As densities across line plots can vary greatly, the sizes of the feature sets also differ heavily. Because the Jaccard index and Sørensen-Dice coefficient are susceptible to variations in the set size, their values tend to be lower when comparing sets of vastly different sizes, even if one is a subset of the other. Therefore, for line-based density plots the overlap coefficient [67] seems to be more suitable:

\[
\text{overlap}(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)}
\]

It is insensitive to differences in set sizes as the denominator is solely determined by the smaller set. In a line-based density plot, the feature set of a bin in a low-density region (with dispersed lines) is typically a subset of the feature set of a bin in high-density areas. Because lines passing through a lower-density bin are likely to also pass through a higher-density bin, where the lines are more aggregated. In this case, we expect the two sets to have a high degree of similarity. The overlap coefficient accurately reflects this phenomenon. A more detailed comparison is available in Section 2 of the supplementary material.

We use average linkage hierarchical clustering [15] as it can be used with arbitrary distance metrics, and the number of clusters does not have to be specified in advance. Additionally, the user can interactively divide the clusters to analyze the data further. We did not use well-established partitioning clustering methods such as PAM to identify the clusters since, here, the number of clusters has to be known in advance.

4.4 Cluster Assignment

In the previous step, we grouped the sampled bins into clusters based on their similarity. However, the bins that were not part of the sample have not been attributed to a cluster as of now. To assign clusters to these bins, we have to calculate the similarity between clusters and bins, which is not easily possible using the above-mentioned overlap measure. Since only the similarity between bins can be calculated, we would have to calculate the overlap coefficient similarity of an unassigned bin...
to all bins of each cluster, which is time-consuming. To solve this, we instead compute a mean feature vector \( M(C) \) for each cluster \( C \). The feature vector \( V(p) \) of the feature set \( S(p) \) of a bin is a binary vector with entries set to 1 if the identifier \( i \) of the line \( L_i \) is contained in \( S(p) \) and 0 otherwise:

\[
V(p) = (v_p^0, v_p^1, \ldots, v_p^n), \quad v_p^i = \begin{cases} 1, & i \in S(p) \\ 0, & \text{otherwise}. \end{cases} \tag{3}
\]

The mean feature vector of a cluster \( C \) is defined as the element-wise sum of the binary feature vectors \( V(p) \) of all its bins \( p \), divided by the number of bins in \( C \):

\[
M(C) = \frac{1}{|C|} \sum_{p \in C} V(p) \tag{4}
\]

To assign a cluster to each bin, we identify the most similar cluster \( C \) for this bin \( p \). To calculate the distance between the bin’s feature vector \( V(p) \) and the mean feature vector \( M(C) \) of the cluster \( C \), we use the Euclidean distance as a similarity measure. Note that when calculating the distance, we only consider the position where \( v_p^i \) is 1:

\[
D(p, C) = \sum_{i \in S(p)} (1 - M(C)_i)^2 \tag{5}
\]

This is similar to our rationale for choosing the overlap coefficient because if we consider all \( v_p^i \), it would also count the lines not touching the bin, which leads to an unreasonable increase in the distance.

### 4.5 Cluster Colorization

In the next step, we assign colors to clusters so that these colors visually represent the similarities among them – bins of similar clusters should be colored similarly, while dissimilar clusters should be colored differently. Thus, a bin’s color solely depends on the cluster it is assigned to and does not, e.g., reflect the orientation of the lines passing through it. A bin’s color should visually communicate both the cluster and the density of the bin. Changing a bin’s cluster assignment as follows:

\[
d_{ij} = \begin{cases} |	heta_i - \theta_j|, & \theta_i - \theta_j \leq \pi \\ 2\pi - |	heta_i - \theta_j|, & \theta_i - \theta_j > \pi \end{cases} \tag{7}
\]

Therefore, for the unit circle, \( \delta(x_{ij}) \in [0, \pi] \). We use the partial derivative \( \frac{\partial^2 S}{\partial \theta^2} \) (details see the Section 1 of the supplementary materials) of the stress \( S \) with respect to the angle \( \theta_i \) of a point \( k \) for allowing gradient descent. By doing so we optimize the angle of each point and this way determine the hue value corresponding to each cluster center.

### 4.6 Cluster-based Line Filtering

Our coloring method identifies the main trends and highlights them through colorization. While being able to analyze individual trends in isolation and identifying the lines contributing to them is an important task, traditional density plots and our colorization aggregate lines in bins and thus do not directly provide a correspondence between trends in the density plot and individual lines. Therefore, we have to assign each line to a cluster of the density plot. To determine the cluster of a line, we focus on two aspects: grouping the bins of each cluster and summing up their weights. Counting the bins of each cluster that a line passes through already gives us a good intuition about which cluster may describe the line best. However, the perceived trends mainly depend on high-density regions, so bins of these regions should be more influential. Inspired by density-based edge clustering [36] and representative line selection [81], we, therefore, also consider the density of the bins a line passes through. Bins are expressed as:

\[
C_k^i = \{ p_j \mid p_j \in C_k \land i \in S(p_j) \} \tag{8}
\]

where \( C_k^i \) represents the set of bins of the \( k \)-th cluster \( C_k \) that line \( i \) passes through. Then, we compute the sum of densities of the bins in the set \( C_k^i \) as follows:

\[
W_k^i = \sum_{p_j \in C_k^i} D(p_j) \tag{9}
\]

where \( D(p_j) \) is the density of bin \( p_j \). Thus, we assign line \( i \) to the cluster \( C_k \) with the greatest summed-up weight \( W_k^i \). This allows users to analyze the details of the lines matched to each cluster.

## 5 Results

To demonstrate the effectiveness of our approach and compare it to traditional coloring schemes of density plots. We selected four datasets: two common time-series datasets that have been used in [44, 81] and two trajectory datasets: Hellenic Trench AIS data [18] and taxi rides in Beijing [76, 77] (featured in the Fig. 1).

**Temperature data.** This dataset contains 293,175 weekly maximum temperature values of 6187 time series obtained from stations across the United States [1]. The line plot in Fig. 4a is highly convoluted and offers no insight into the underlying trends in the data. The conventional density plot in Fig. 4b visually separates high from low-density regions, but there are ambiguities about how individual trends continue after crossing, e.g., in the region marked “A” on the right. In addition, the parts of the trends circled as “B” show zigzag features and align with the ambiguous continuation problem shown in Fig. 2b. Our coloring method in Fig. 4c helps to reduce these ambiguities and reveals patterns that are not easily visible otherwise. In particular, the blue cluster is visually separated from other trends, indicating areas with consistently high temperatures, which makes region “B” more clearly stand out. Meanwhile, the purple and green trends intersect at two positions (marked as “A” in Fig. 4c), revealing areas that experience sudden temperature changes, making them more anomalous at certain times. The line counts for these clusters were 2180 (blue, Fig. 4d), 793 (green, Fig. 4e), and 742 (purple, Fig. 4d). This demonstrates that there are indeed many lines that follow the trend of these clusters. The lines in the blue cluster remain at higher temperatures, while lines of the green...
Fig. 4: Time series of temperature values (6187 lines) visualized using (a) line-based visualization, (b) line–based density plot, and (c) our line-based density plot colorization scheme. (d) - (f): Individual trends selected using our interactive tool.

Fig. 5: Daily closing prices of 4393 stock at the New York Stock Exchange from 2005 - 2017 visualized using (a) line-based visualization, (b) line-based density plot, and (c) our line-based density plot colorization scheme. (d) - (f): Individual trends selected using our interactive tool.

Fig. 6: Hellenic Trench AIS data [18]: (a) density-based visualization of randomly selected 10,000 trajectories with 1,048,575 points. A “U”-shaped trend seems to dominate the plot; (b) The colored density-based plot instead highlights several high-density trends that intersect; (c) The lines associated with the green cluster show a converging pattern of the cluster after vessels passed the channel (from west to east) between two islands.
and purple clusters cross at certain positions, supporting our previous conclusions drawn from the colored line-based density plot.

Stock market data. In the upper row of Fig. 5, we show 552,559 daily closing prices of 4393 stocks (with prices between 0 and 100; we removed stocks with only one close price) at the New York Stock Exchange [2] from 2005 to 2017. Although the raw line-based visualization in Fig. 5a is cluttered, it is still possible to observe that there are many lines at the bottom of the graph. The pattern (highlighted as “A”) at the bottom of the density plot in Fig. 5b aligns with this observation. However, there is another high-density pattern, which is shorter and concentrated towards the end of the graph (highlighted as “B”), that only appears in the density plot. These two disconnected high-density patterns show the issue of disconnected clusters we presented in Fig. 2c in a real-world dataset. Our colorization in Fig. 5c illustrates that the clusters marked “A” and “B” are actually disconnected. The cluster “B” does not start from the beginning of the time period but instead starts from in between. This is because many stocks were newly listed after 2010 and did not have corresponding records before that time (see Fig. 5d). After matching lines to the three clusters in the colored line-based density plot, the lines of the individual clusters are shown in the lower row of Fig. 5. The green cluster represents the main trends and contains to more than half of the total lines (Fig. 5e). It represents a large number of stocks fluctuating at lower prices. The red cluster contains fewer lines, but the prices of those stocks are relatively high, even though they are more scattered (Fig. 5f).

Ship trajectories. The Hellenic Trench AIS dataset [18], containing more than 170,000 vessel trajectories, was originally collected to analyze the impact of ship routes on the survival of Mediterranean sperm whales [19]. At first, we cleaned the dataset by removing trajectories crossing land, likely introduced by GPS recording errors. Due to the memory limitation of the browser, we randomly selected 10,000 trajectories with 1,048,575 points to generate the density plot and our colorized version. The density plot in Fig. 6a highlights several intersecting high-density trends. Visually, a dominant “U”-shaped trend (marked as “A” in Fig. 6a) begins west of the Peloponnese and continues eastwards of Athens. Similar to Fig. 2a, based on the density map alone, it is unclear if the trend is illusory or not. After coloring the density plot by our method (Fig. 6b), the western and eastern parts of the “U”-shaped trend are colored blue and green. Thus they are likely separate patterns. Through our line assignment method, 2604 lines were assigned to the green cluster, as shown in Fig. 6c. The lines of the green cluster are almost on the same route after entering the channel but beforehand come from various directions. Although there are some trajectories that match the visual “U”-shaped trend, such lines are not as dense as indicated in the original density map. Our method succeeds in finding lines that match other clusters and is able to decompose them into several clusters of line bundles. Another point worth noting is that the pixels in the lower left corner of Fig. 6b (marked as “B”) look like the lines are headed in one direction, but they have been grouped into different clusters. This reflects a limitation of our approach: it is not able to, e.g., measure parallelism or other geometric features of lines.

Taxi trajectories. The Beijing taxi trajectory dataset [76, 77] contains GPS trajectories of 10,357 taxis on February 2-8, 2008. As this dataset is very large, loading it in our tool exceeded the browser’s memory limitation. Thus, we sampled 6502 trajectories of 500 taxis with 734,967 individual time points to generate Fig. 1. The line plot in Fig. 1a shows severe overplotting, hiding most of the dataset’s structure. However, although the line-based density plot in Fig. 1b reveals the road network of Beijing, it fails to display frequent taxi routes, e.g., it is ambiguous if the taxis circle around the city center or if they drive outwards or inwards from there. Our method allows to identify different clusters. The lines associated with the green and purple clusters in Fig. 1c are shown in Fig. 7. While the green cluster represents taxis driving along the airport highway, where many taxis exit at the toll booths to reach different destinations, the purple cluster represents taxis traveling within the eastern part of Beijing. This can be explained by the presence of a large residential area east of Beijing. Similar patterns can be observed for the other clusters. Taxis are unlikely to circle the city center but are likely to drive in one part of the city.

Fig. 7: Filtered lines from two clusters of Fig. 1c.

Fig. 8: Comparison between colorized line-based density plots and line-based clustering for the scenarios of Fig. 2.

5.1 Comparison to Line Clustering Approaches

We use the three synthetic datasets shown in Fig. 2 to illustrate the results of coloring lines by clusters. The dataset used for Fig. 2a contains 400 lines representing the actual pattern and, additionally, 100 lines of background noise. Figs. 2b and 2c contain 200 lines per pattern.

Existing methods treat lines as basic data units and focus on line-based analysis to find trends in the data [17] using clustering. To assign colors, the lines are clustered using a similarity measure tailored to line data. While these approaches work well for lines with distinct line bundles, they do not work well for noisy data. To highlight the limitations of such methods, we use the state-of-the-art line-based method proposed by Ferstl et al. [17] and compare their results to ours. Their approach involves embedding lines into a higher-dimensional vector space, followed by clustering in a lower-dimensional Euclidean space after using PCA.

To compare our approach to the line-based method, we apply both methods to the examples shown in Fig. 2. While the line-based method fails to distinguish the line bundles in two of the three cases due to a less separable data distribution in low-dimensional space, our approach effectively reveals the underlying structure. For instance, in Fig. 8a, our method colors the pixels with two colors, showing that the pattern is not a continuous trend (marked as “A”, “B”, and “C”) but is composed of different individual patterns. In contrast, the result of the line-based approach in Fig. 8b cannot reveal the actual components of the trend. Similarly, in Fig. 8c, our method provides a clear view of the trends of the two line bundles, achieving results that are similar to the line-based method (Fig. 8d). Finally, our approach successfully divides the pixels of the disconnected clusters (Fig. 8e), whereas the line-based method (Fig. 8f) fails to provide valuable information in this case. Section 5 of
We conducted a browser-based online user study with university students and Ph.D. candidates. A total of ten line datasets were included, all presented in the form of line-based density plots. Participants had to identify trends in line-based density plots and complete three tasks for each dataset. All plots used in the user study and the trends they are composed of are available in Section 3 of the supplementary material.

6.1 Experimental Design

We conducted a browser-based online user study with university students and Ph.D. candidates. A total of ten line datasets were included, all presented in the form of line-based density plots. Participants had to identify trends in line-based density plots and complete three tasks for each dataset. All plots used in the user study and the trends they are composed of are available in Section 3 of the supplementary material.

6.2 Implementation and Interaction

Implementation. Results were obtained using an Apple Silicon M1 processor with 16GB RAM. Our interactive system (https://color-line-density-plot.github.io/) is written in Javascript and does not utilize the GPU. We used Firefox for the runtime tests. The time necessary to calculate the overlap coefficients is dependent on the size of the dataset, resulting in longer computation times for larger sets. For 10,000 lines, our hierarchical tree construction needs about 10s. Classifying all bins per division of clusters took less than 4s. The processing speed improves significantly with decreasing cluster sizes. In summary, our method can complete pre-processing (including threshold filtering, sampling, and hierarchical tree construction) for all datasets presented in this paper within 20s. Interactions to separate clusters took less than 4s, mostly between 1s and 2s.

Density threshold and sampling rate. Section 4.2 outlines the sampling strategy that requires setting density threshold and sampling rate. To enhance the intuitiveness of the minimum threshold, bins below the threshold are hidden using a corresponding slider. The user is informed when the selected bins exceed the processing capacity.

Cluster operations. Users can set the number of clusters as well as interactively separate clusters. After setting a number, clusters will be divided accordingly, based on the distance between the nodes of the hierarchical tree. Clicking on a cluster allows the user to split it into two sub-clusters. The feature vectors of these sub-clusters will then be mapped using circular MDS to find appropriate hue values. Users can select cluster colors that are fixed in the hue circle. Then, during gradient descent, only unfixed points are moved.

Hue adjusting and harmonization. In Section 4.5, we describe our use of multidimensional scaling (MDS) to ensure that the distance relationships between hues within each cluster correspond to their distance relationship in high-dimensional data space. Nonetheless, users can modify the hue by dragging the dot representing the cluster on the hue wheel. Additionally, we provide users with eight harmonic templates, as defined by Cohen-Or et al. [12]. Once a template is selected, the hues for each cluster will be mapped to the corresponding area, with the movement of the dots being constrained accordingly.

6 Motivational User Study

Here, we present the findings of a small motivational user study we performed to understand the ability of users to identify and trace trends in line-based density plots. We explored whether users have issues separating trends of lines. Additionally, we investigated if users are even aware of potential ambiguities in given plots. We did not include a comparison between our proposed colorization method and traditional density plots. We expected participants to perceive different trends overall, but as long as the line-based density plot remains visually similar, even if it is generated based on different underlying data, their perceptions should remain unchanged. We posed the following two hypotheses:

H1: There are different understandings when participants perceive trends in ambiguous line-based density plots.

H2: The participants are mostly unaware of the ambiguities in line-based density plots.

6.1 Experimental Design

We conducted a browser-based online user study with university students and Ph.D. candidates. A total of ten line datasets were included, all presented in the form of line-based density plots. Participants had to identify trends in line-based density plots and complete three tasks for each dataset. All plots used in the user study and the trends they are composed of are available in Section 3 of the supplementary material.

Tasks:

- Tracing task. Participants were asked to trace each dominant trend they identified.
- Counting task. Participants were asked to count the number of trends they perceived for each line-based density plot.

Ambiguous pattern generation. We generated ten line datasets in total based on the three types of ambiguities introduced in Section 3.2. For each kind of ambiguous pattern, we provided multiple examples with varying degrees of ambiguity or different numbers of trends. In the first three datasets, we used illusory patterns, similar to Fig. 2a. The line-based density plot of the first dataset was composed of a horizontal trend superimposed on noise, which is similar to the right trends in the figure. For the following two datasets, we combined trends like the ones on the left in the top row of the figure. We increased the fan out in the center of each individual trend. Thus the plots were getting progressively less ambiguous, as seen in Fig. 10b. As the second kind of ambiguous pattern, we added ambiguous continuations to datasets 4 to 6. The combined patterns from Fig. 2b, both the upper left and the upper right composition, are used for datasets 4 and 5. In addition, we added another line dataset whose density plot was composed of three individual trends (dataset 6). The third kind of pattern showed disconnected clusters. In the last four datasets, the density plots contained potentially disconnected clusters. Datasets 7 and 8 were composed of distinct patterns. On one hand, dataset 7 consisted of a continuous trend with a central fan-out region superimposed on low-density noise, as shown on the right side of the upper row in Fig. 2c. On the other hand, dataset 8 comprised two independent trends, which are illustrated in Fig. 10d. To decrease the ambiguity of the pattern, we increased the horizontal length of the high-density regions of the two independent trends to create datasets 9 and 10. This reduced the size of the central overlapping region as shown in Fig. 10e.

Experiment organization. We displayed the line-based density plots of the datasets colored using a multi-hue color map, one dataset per page, on our survey website. All participants had to fulfill the three tasks for each of the ten datasets in the same order. For the datasets with illusory patterns (datasets 1 to 3), because we were interested in whether participants could tell if multiple trends were combined, we asked them to trace only the main trends. For the datasets with ambiguous continuation (datasets 4 to 6), we asked participants to trace the individual trends they perceived with different color. For the disconnected clusters (datasets 7 to 10), we asked participants to mark the trends they thought were connected with the same color of strokes, and if they thought they were disconnected, use different colors.

Participants. In total, we collected inputs from 25 participants who completed our user study effectively. Three other participants were excluded as they either did not trace any trends or entered numbers of perceived trends that vastly differed from the trends they actually drew. We gathered the traces for all ten tasks drawn by each participant. The participants’ age was, on average, 26 years with a standard deviation of 4. The gender distribution among our participants was 56% male, 36% female, and 8% preferring not to tell their gender.

Procedure. Each participant went through the following procedure: (1) completing a short questionnaire with questions about age, gender, and familiarity with line charts and density plots; (2) viewing an introduction about the tasks based on an already marked density plot; and (3) performing the tasks for each dataset.

<table>
<thead>
<tr>
<th>Ambiguous patterns</th>
<th>Certainty task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illusory patterns</td>
<td>100% certain</td>
</tr>
<tr>
<td>Ambiguous patterns</td>
<td>90% certain</td>
</tr>
<tr>
<td>Disconnected clusters</td>
<td>80% certain</td>
</tr>
</tbody>
</table>
6.2 Experimental Results

Fig. 9 illustrates the participants’ responses using a Likert scale to the questions about their certainty regarding the number of trends and the tracings of the trends. Fig. 10 shows the distribution of the number of trends the participants perceived for each line-based density plot.

**Illusory patterns.** Most participants perceived a central horizontal trend in the density plots of the illusory patterns shown in Fig. 10a. The first dataset, consisting of only a horizontal trend superimposed by noise, was correctly identified by all participants and traced using a horizontal line. For all datasets with illusory patterns, more than 70% participants stated that they were “Moderately certain” or “Extremely certain” about the number of trends and their tracings. However, increasing the fan-out region of the combined trends in the illusory pattern, as shown in Fig. 10b, decreased the participants’ certainty about how many trends they perceived and how they traced them. Fig. 9 shows a decrease in the portion of participants being “Extremely certain” about the number of trends they identified and how they traced them from the first dataset to the third. Additionally, the portion of participants perceiving at least two trends increased from 0% to 16%, as shown in Fig. 10. This indicates that participants were increasingly unsure about the trends in the density plots. Still, even though datasets 2 and 3 were created from data containing multiple trends, participants mostly perceived only a single trend for all datasets (92% and 84%, respectively), which supports **H2**.

**Ambiguous continuation.** All participants correctly identified, with moderate to high certainty, that the combined line-based density plots of the individual ambiguous pattern shown in Fig. 2b are composed of two trends. For dataset 6, consisting of three trends, all participants except for one correctly identified the number of trends. In the tracing task, 88% of the participants identified the same trends in both datasets 4 and 5. 60% of the participants traced the combination of two “U”-shaped trends in both plots. While 28% of the participants traced crossing trends for both plots in an “X” shape. For datasets 4 and 5, whose density plots are almost identical, we expect participants to perceive them ambiguously. For dataset 4, 16 participants identified a “U”-shaped pattern and 9 an “X”-shaped. For dataset 5, 17 participants traced a “U” and 8 an “X”. For dataset 6, which contains a combination of three trends, the participants’ understandings are more diverse. We only offered the ability to trace a single trend in the data. Participants were aware that there might be multiple possible trends in the plot but drew only the most dominant one. Therefore, the results do not show whether participants potentially perceived multiple patterns. The distribution of participants among both identified trends indicates that there was no consensus on which trend was dominantly perceived. We conclude that this supports **H1** with respect to the group of participants, but not for an individual participant.

**Disconnected clusters.** While participants largely agreed on the number of perceived trends in the previous patterns, they were torn about whether the disconnected clusters in Fig. 2c contained one or two trends. For dataset 7, 56% and for dataset 10, 64% of the participants indicated that the trends were independent, like in the left example of Fig. 2c. The remaining portion of participants highlighted just one continuous trend, as shown on the right. Thus participants did not agree, while over 80% of them were at least moderately certain about the number of patterns they identified and their tracings. Fig. 10c shows that the disagreement of the participants regarding the number of trends for datasets 7-10. As the density plots of datasets 7 and 8 are almost identical but contain different trends, we expect participants to give similar answers. For dataset 7, 44% of the participants answered one trend, and 56% of the participants answered two. While for dataset 8, it is 40% and 60%, respectively. For datasets 9 and 10, the participation response percentages are still similar, although they are less ambiguous. This further supports **H1**.

6.3 Experimental Findings

While our user study was meant to exemplify the ambiguity in the user’s understanding of line-based density plots, it is not a large-scale study with extensive possibilities for statistical evaluations. Nevertheless, it supports our hypothesis **H1** and **H2** and shows trends in line-based density plots are perceived ambiguously. While participants can correctly identify the number of trends for the patterns that showed ambiguous continuation, they disagreed about the flow of the trends. Our results indicate that although there is a great deal of disagreement among the participants regarding the underlying trends, the individual participants seemed to be fairly certain about the number of trends and their tracing. Thus it is important to support users in identifying individual trends and avoiding wrong perceptions.

7 Discussion and Conclusion

In this paper, we presented a novel coloring method for line-based density plots. Although our approach is an image-based analysis of line data, it could be applied to density plots of parallel coordinates or even other types of data. In contrast to the in Section 2 introduced line-based methods, which group similar lines by colorization of the lines, abstract them with statistical information of ensembles, or combine them by altering their paths using edge-bundling, we aggregate the individual line into line-based density plots and enrich them by clustering similar regions in the image space.

**Future works and limitations.** The basic unit of data for our presented method is pixels, which means the higher the resolution of the visualization, the more computationally demanding it is. However, it is unnecessary to perform binning for each pixel. Different bin sizes can be used in different areas depending on the distribution of the data. For example, a larger bin size can be used for unimportant areas. So the plot could be divided into differently-sized bins to reduce the number of bins. Therefore, the next step would be to explore automatic methods for finding the best binning strategy. In subsequent works, we also want to reduce the number of clustering errors by investigating other sampling strategies.

Edge bundling is a widely used approach to address the clutter in line-based visualizations. Recent methods address the issue of ambiguities introduced by edge bundling. Our method could also be applied to disambiguate further the results generated by edge bundling methods.

With an increasing number of clusters, the angle between assigned colors on the hue wheel decreases, which reduces their distinguishability. Already for 7 different colors (see Fig. 6b), the colors start to be barely distinguished. Furthermore, perception depends on cluster sizes, spacing of hues, and positional relationships between clusters. Conducting a user study could shed light on these aspects.

Our small-scale motivational study was only concerned with showing whether and to what extent the perception of trends in line-based density plots is ambiguous. It encompassed only university students and Ph.D. candidates. But it underpinned our hypothesis that line-based density plots are indeed perceived ambiguously by users. We investigate only whether the user’s understanding of line-based density plots is ambiguous. Therefore, we did not include a comparison of line-based coloring methods and our method in the user study. A further study should have a larger and more diverse population, including perception and understanding of our results.

Currently, we summarize three types of ambiguities in line-based density plots, but it is possible that additional types also exist. In future work, we intend to develop a more comprehensive taxonomy to categorize the ambiguities found in line-based density plots.
ACKNOWLEDGMENTS
This work was supported in part by Deutsche Forschungsgemeinschaft (DFG) Project 410883423, Project 251634672 – TRR 161 “Quantitative methods for visual computing”, KE 740/17-2 of the FOR2111 “Questions at the Interface”, the National Key R&D Program of China (2022ZD0160805), NSF China (No.62132017, 621-41217), and the Shandong Provincial Natural Science Foundation (No. ZQ2022QJ32).

REFERENCES
[34] A. Mayorga and M. Gleicher. Splatterplots: Overcoming overdraw in scatter plots. IEEE Transactions on Visualization and Computer Graphics,