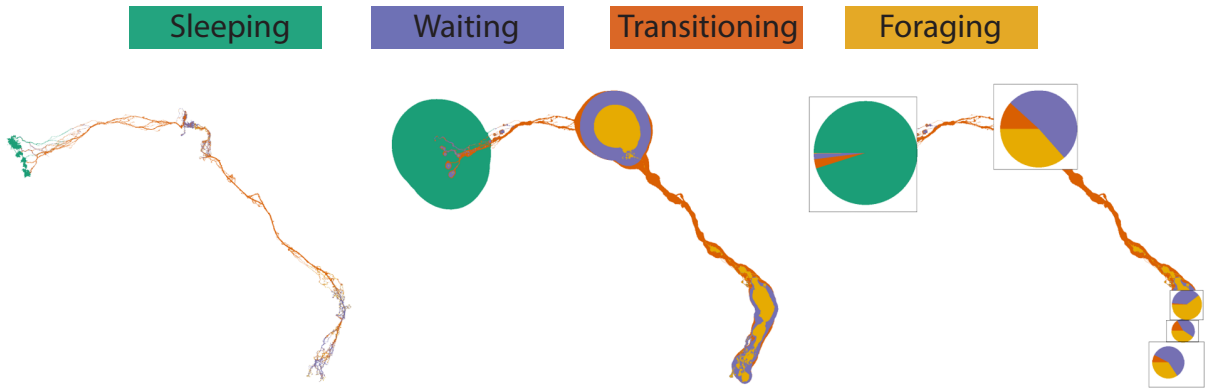
# Pixel Wise Pie Charts: Placement of Data Points Visualizing Spatial Value Distributions

Halldor Janetzko[1], Manuel Stein[2]

[1]University of Zurich, Switzerland
[2]University of Konstanz, Germany




(a) original movement data (b) circular point placement (c) pixel-based pie charts

Figure 1: Comparison of three different visualizations of the same animal movement in Kenya. The left figure shows the raw information by colored points representing contextual, ecological data (green means sleeping, purple waiting, orange transition, and yellow foraging). The center figure is created by applying circular pixel placement showing the dense point areas. The right figure results from applying the novel angular, pie-chart placement algorithm presented in this paper to selected regions.

**Abstract**

*Handling overplotted points in spatial pixel-based visualizations is usually done either by aggregation or sampling. The third rather unusual alternative is replacing data items untangling overplotted regions. Depending on the placement algorithm visual artifacts will be created based on an optimization between overplotting, display usage, and conservation. In this paper, we exploit the replacement and create pixel-based pie-charts reflecting the numerical distribution of spatially referenced data points. With this novel approach, we bridge the gap between the overview and detail phase in Shneiderman's Information Seeking mantra. Our presented algorithm is informally evaluated with a biological case study on baboon movement.*

## 1. Introduction

Using a two-dimensional coordinate system encoding data features is a very common approach in information visualization to reveal dependencies and correlations. The transformation from data space to screen coordinates should reveal patterns by spatial groupings or topological properties in the visualization. Furthermore, information visualization techniques are designed to represent similar data properties by similar visual encodings. With respect to spatial layout in case of scatterplots or dot maps this results in small distances between similar data items. However, as there is not an indefinite amount of screen pixels, too similar data items will be mapped to the same screen coordinate. As a consequence, overplotting occurs in dependency of screen resolution and data value range. In Equation 1, we formalized the relationship of data range of a data set $D$ and number of pixels by determining the minimum data resolution $o_D$ in a dimension $dim$ that can be shown on a screen (assuming linear normalization).

$$o_D(dim) = \frac{|max_D(dim) - min_D(dim)|}{pixelsPerDim(dim)} \quad (1)$$

Overplotting of distinct data points occurs, if these data points have a smaller distance than $o_D$ introduced above. In a more formal way, two non-identical points have to be closer than $o_D(dim)$ in

either the $x$ or $y$ dimension. The set of overplotting pairs of points $O$ is described in Equation 2.

$$O = \{(p,q) | p,q \in D \wedge q \neq p \wedge \exists dim \in DIM : |p_{dim} - q_{dim}| \leq o_D(dim)\} \quad (2)$$

The inherent problem of overplotting is the loss of density information and consequently very dense point clouds may look less populated than spread point clouds. This problem is well-known in visualization and several solutions have been proposed such as statistical sampling methods to reduce the number of points being displayed and decrease the overplotting as well. Nevertheless, overplotting still occurs and density is not easily accessible. Therefore, density aggregation views as kernel density estimations or grid-based approaches can be used to reveal the data distribution. However, single data points are not accessible anymore being a severe drawback when a third dimension needs to be displayed for example by colorization. Previous research also proposed distortion and point placement techniques changing the position of the data point. Moving data points to the nearest empty pixel position resolves any overplotting paying the price of introduced visual artifacts as described in the technique Section 3. In our teaser Figure 1, we show the application of a circular placement algorithm (center) to a spatial data set with overplotting (left). The circular visual artifacts do show the number of points but are highly dependent on the order of point placement. Due to the quadratic increase of pixels covered with increased radius, colored points placed last (yellow, purple, orange, and lastly green in Figure 1) will always look less frequent as they are. Consequently, the design requirements for our visualization design are as follows:

- overplotting free visualization of two-dimensional data points
- visual encoding of thematic information for each data point
- visualization of spatial distribution and ratio of thematic classes
- all data points are visualized and accessible by interaction

In this paper, we contribute a novel placement algorithm satisfying the requirements mentioned above revealing the ratio of thematic value distributions by creating pixel-based pie charts. The result of this new technique can be seen in the right of Figure 1 and is further discussed in Section 4. The resulting visualization contains by algorithmic design no overplotting and each pixel represents at most one data point. Our technique inherently connects a visual overview aggregation with the interactive details-on-demand phase of Shneiderman's Information Seeking mantra as each pixel of a pie charts represents one underlying data point.

## 2. Related Work

Two-dimensional point based visualizations are not a new technique and are used for either spatial phenomena or correlation analyses of two dimensions. A well-known and often cited cartographic example is the dot map visualizing cholera cases in London 1854 designed by John Snow [Sno55]. Scatterplots do offer a large design variety as described in [SG18] by Sarikaya et al. Depending on the task, different design alternatives are more suitable than others. Scatterplots as statistical visualizations and dot maps convey patterns by point distributions and densities. With overplotting being a persistent challenge, Cleveland presented in 1984 for example a glyph-based representation depicting the data density called sun flowers [CM84]. However, additional glyphs may introduce other visual artifacts, such as overplotting of glyphs or binning artifacts. In geographic map design, this technique is very close to proportional symbol maps coming with the same advantages and disadvantages. Visual sedimentation was proposed by Huron et al. in [HVF13] to reduce and abstract the visualized data. Visual sedimentation transforms data points to new geometrical shapes while our technique retains the original data points. A good overview of methods for multidimensional data exploration is provided by Hurter in [Hur15].

Semi-transparency of drawn data points may be used to convey a notion of density and is described in [Wil06]. In most cases using transparency will help analysts to investigate the density distribution. A severe drawback of transparency is the limited readability of resulting density visualizations. Perception-wise, transparency is not correlating linearly to the number of points painted at one single position and may lead to wrong impressions. Setting an optimal transparency value is very challenging and depends highly on the data set and the task. Often, global transparency values are not sufficient because of skewed density distributions.

Density visualizations display aggregated information, but not the raw point data as traditional scatterplots. Consequently, density visualizations cannot show a third value by coloring. Bowman and Azzalini [BA97, BA03] proposed for example smooth contour scatter plots showing overlaps with different shades. Continuous scatter plots invented by Bachthaler and Weiskopf [BW08] derive from the discrete input data a continuous model. Continuous scatter plots do not suffer from overplotting points as they are aggregated when building the underlying model, but the analyst loses the notion of how many data points lead to the visible patterns. Mayorga and Gleicher discuss in [MG13] a technique combining density aggregation with point-based representations in scatterplots. The authors apply perception-based color blending for density surfaces and sampling for visualizing a subset of the original data points.

Relocating data points (and consequently distorting the data space) is the second option when dealing with overplotting. We build on our previous presenting a technique called Generalized Scatter Plots [KHD*09] allowing analysts to interactively control the amount of distortion and circular pixel placement. We proposed further enhancements in [JHM*13] to optimize the placement and increase the visual saliency of location shifts. Another technique actively reducing the amount of overplotting was introduced by Trutschl et al. in [TGC03]. The authors present a SOM-based approach replacing overplotting data points according to the points' similarity. A very simple, row- or column- based pixel placement algorithm is presented by Aris et al. in [AS07]. This approach will introduce visual artifacts resulting from the row- or column-based replacement of data points. Using interactive lenses to enhance the visual representation of interesting regions enables analysts revealing insights. A comprehensive overview of techniques and usages of smart lenses is given by Tominski et al. in [TGK*14]. The additional information presented within the smart lenses usually overplots the visualized data below. Chen et al. propose in [CCM*14]

hierarchical sampling decreasing overplotting while retaining class ratios. This technique does not remove overplotting but increases the readability of scatterplots by visualizing a well-chosen selection of data points. Providing proper interaction methods as proposed by Hurter et al. in [HTCT14] enables analysts dealing with large amounts of overplotting in point clouds.

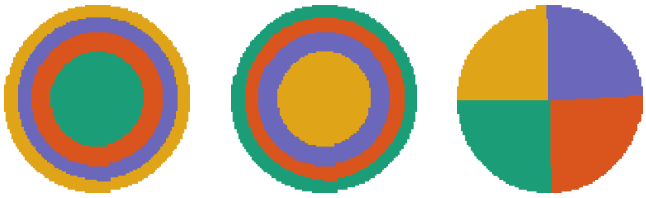
## 3. Technique

The basic and very abstract principle of point replacement algorithms avoiding overplotting can be described as shown in Algorithm 1. Depending on the algorithmic definition of closeness marked by • different shapes will result. In our case, we choose a quadratic norm leading to data points being moved in a circular manner.

**Input:** Ordered list of all data items
**foreach** *DataPoint cur* **in** *allDataPoints* **do**
  **if** *cur* can be placed at original position **then**
    assign original position to *cur*
  **else**
•   retrieve closest empty pixel for original location of *cur*
    assign new pixel position to *cur*
  **end**
**end**

**Algorithm 1:** Abstract pixel-based algorithm for overplotting-free point placement

We briefly mentioned in the introduction the problem of comparing value distributions within a circle. We exemplify this problem in Figure 2 by an artificial data set of 6000 data points divided in four equally-sized groups all located at the same coordinate. Although all three visualizations show the same data set and are free of any overplotting, comparing the ratio of values is not easy for the first two cases.



(a) ascending (b) descending (c) angular

Figure 2: Comparison of values ratios hindered or enabled by different ordering methods during point placement.

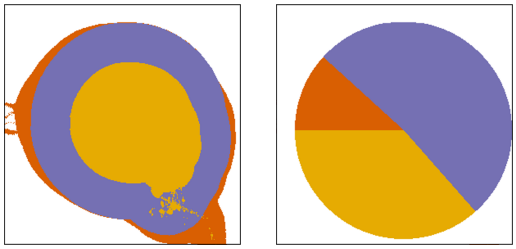
The highly task-dependent grouping of data points to reveal meaningful insights is done by human analysts as an initial step. The user selected regions (black rectangles in Figure 1c) are the input for our pie-wise pixel placement algorithm. The basic idea is to employ the output of the circular placement algorithm for each region and arrange all data points in an angular fashion. Consequently, the pie-wise placement algorithm consists of multiple consecutive steps.

At first, we iterate over all selected regions and determine for each region all contained data items. For each of the regions, we compute the centroid and assign to all data items this computed location. Consequently, we collapse all points of one region to the same location and introduce as an intermediate step high amounts of overplotting. We remove the overplotting by applying the circular placement algorithm presented in [KHD*09]. The final post-processing step will resort all points of each filled circle in-situ within each circle. The resorting is done by angular assignments of ordered (by angle to centroid) locations to the ordered (by value) data points.

## 4. Application

We demonstrate the usefulness of our approach by applying our technique to the analysis of animal movement data. In particular, we focus on the analysis of different activity phases of baboons. The data for this case study is resulting from previous work of Strandburg-Peshkin et al. [SPFCC15]. The data consists of 26 tracked baboons (81 % of the complete baboon group) in Kenya over the course of two weeks as well as meta data for every baboon. Positional coordinates are available once per second from 6am to 6pm. Strandburg-Peshkin et al. analyzed how individual baboons decide which (sub-)group to follow. Furthermore, they investigated possible prediction models for future baboon movement. However, there was no in-depth analysis of the various forms of activities as well as the spatial distribution of activities.

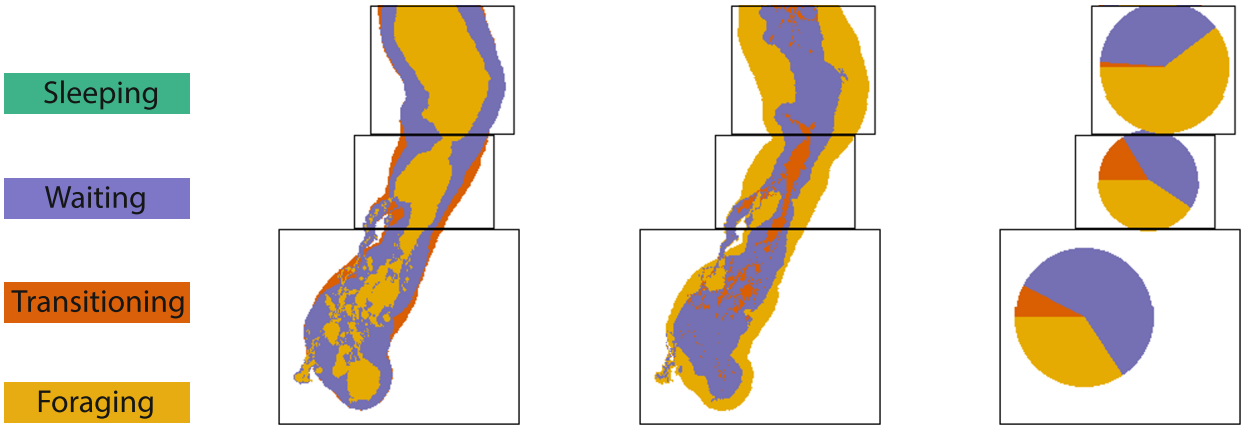
In this case study, we visualized the baboon movement with state-of-the-art methods and our novel placement algorithm as depicted in Figure 1. We differentiate between four kinds of activities. The baboons always **sleep** (green color) in the same area every night. After waking up, they **transition** (orange color) as group in a line-like formation until they reach an area for **foraging** (yellow color). In order to make foraging more efficient, baboons are known for moving slowly and spreading out while looking for food. While other baboons might still be foraging, some baboons also just **wait** (purple color) for the other baboons to join the collective group movement again.



(a) circular (b) angular

Figure 3: Enlarged visualization for a selected area of interest (corresponding to right topmost selection in Figure 1). The ratios of activities is better salient in the angular point placement.

In Figure 1 (a), we see the original movement of 200,000 data points mapped accurately with respect to spatial location corresponding to movement of one day from sunrise to noon. Due to overplotting, it is impossible to get an adequate impression of time




(a) activities (b) foraging placed first (c) transitioning placed first (d) angular

Figure 4: We enlarged the foraging grounds in the southeast of the analyzed baboon data set. Interesting is the missing transition (no orange) activity in the top selected region. One possible explanation for the missing transition is the richness of food sources in this area.

spent at a location and the spatial distribution of collective activities. Figure 1 (b) shows one state-of-the-art solution for this overplotting problem employing a pixel-based circular point placement method. Comparing the different activity frequencies is hard because of the circular layout as described in Section 3. Our proposed method (Figure 1 (c)) improves the visual perception of the occurring activity types. A side-by-side comparison for a specific dense region (right topmost selection in 1 (b)) is given in Figure 3. The wrong impression of frequencies – purple points (waiting) seem to be less frequent than they actually are – is obvious when placing both visualizations side by side. The alignment of the points in Figure 3 (a) give the impression that foraging (yellow) is dominating in this area while Figure 3 (b) clearly illustrates that the baboons were mostly waiting.

The importance of ordering during point placement is illustrated in Figure 4. By focusing on the main areas of foraging (southeast of baboon movement in Figure 1) we can extract insights of the food availability. As focusing on foraging, we place the data points corresponding to foraging activity first as shown in Figure 4 (b). An in-depth inspection reveals that there seems to be less transitioning in the top area of interest. Knowing of the perceptual bias resulting from placement order, we change in Figure 4 (c) the order of point placement and place transition activities first. The ratios of activities revealed by the pixel-based pie chart presented in Figure 4 (d) shows the highest relative (and absolute) number of foraging activities in this northern region.

## 5. Discussion

Pie charts are not the best technique to show ratios of different classes. The respective angular position of classes has an influence of the perceived angle. Furthermore, human perception is not meant to judge angles but rather length or size. Nevertheless, we relocated data points creating pie charts, because of the visual similarity of the circular point placement and pie charts. In future work, we need to compare the effectiveness of pie charts with other statistical visualizations as for example bar charts. In these comparisons, we need to assess the introduced placement error as well.

Ensuring an overplotting-free visualization is always coming with a certain price. Aggregations will lead to a loss of detail information and distortions will introduce visual artifacts. Point placement methods have to be seen in the context of the explorative data analysis process as one of several techniques revealing specific data aspects. Point placement changes the location of points but reveals dense areas and value distributions in these dense regions.

The order of point placement is usually crucial for the effectiveness of placement algorithms. The effectiveness of our presented pixel-based pie placement algorithm is not that dependent on the order of data values. We exemplified for our technique that comparing the frequency of data values is not as affected by order.

The inherent challenge of selecting regions to create statistical pie charts is the size of the selection. Dependent on the size and location there will be spatial patterns revealed or hidden. In geography, a closely related phenomenon is known as the *modifiable areal unit problem* describing statistical biases introduced by different aggregations. We consequently enabled the analyst to select the regions of interest reflecting the respective analysis task. Nevertheless, we envision a semi-automatic proposal of regions for creating meaningful selections as future work.

## 6. Conclusion

We presented a novel pixel-based technique removing overplotting in two-dimensional point visualizations enabling ratio comparisons. The statistical visualizations are created from the original data points directly reflecting the number of data points. Furthermore, the connection of a pie chart with the underlying data items is possible as each pixel of the pie chart corresponds to one data item enabling details-on-demand by hovering for example. We see this technique bridging the gap between overview and detail visualizations.

## References


[AS07] ARIS A., SHNEIDERMAN B.: Designing semantic substrates for visual network exploration. *Information Visualization 6*, 4 (2007), 281–300. 2

[BA97] BOWMAN A. W., AZZALINI A.: *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, vol. 18. OUP Oxford, 1997. 2

[BA03] BOWMAN A. W., AZZALINI A.: Computational aspects of non-parametric smoothing with illustrations from the sm library. *Computational statistics & data analysis 42*, 4 (2003), 545–560. 2

[BW08] BACHTHALER S., WEISKOPF D.: Continuous scatterplots. *IEEE transactions on visualization and computer graphics 14*, 6 (2008), 1428–1435. 2

[CCM*14] CHEN H., CHEN W., MEI H., LIU Z., ZHOU K., CHEN W., GU W., MA K.-L.: Visual abstraction and exploration of multi-class scatterplots. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1683–1692. 2

[CM84] CLEVELAND W. S., MCGILL R.: The many faces of a scatterplot. *Journal of the American Statistical Association 79*, 388 (1984), 807–822. 2

[HTCT14] HURTER C., TAYLOR R., CARPENDALE S., TELEA A.: Color tunneling: Interactive exploration and selection in volumetric datasets. In *Visualization Symposium (PacificVis), 2014 IEEE Pacific* (2014), IEEE, pp. 225–232. 3

[Hur15] HURTER C.: Image-based visualization: interactive multidimensional data exploration. *Synthesis Lectures on Visualization 3*, 2 (2015), 1–127. 2

[HVF13] HURON S., VUILLEMOT R., FEKETE J.-D.: Visual sedimentation. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (Dec. 2013), 2446–2455. 2

[JHM*13] JANETZKO H., HAO M. C., MITTELSTÄDT S., DAYAL U., KEIM D.: Enhancing scatter plots using ellipsoid pixel placement and shading. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on* (2013), IEEE, pp. 1522–1531. 2

[KHD*09] KEIM D. A., HAO M. C., DAYAL U., JANETZKO H., BAK P.: Generalized Scatter Plots. *Information Visualization Journal (IVS)* (2009). 2, 3

[MG13] MAYORGA A., GLEICHER M.: Splatterplots: Overcoming overdraw in scatter plots. *IEEE transactions on visualization and computer graphics 19*, 9 (2013), 1526–1538. 2

[SG18] SARIKAYA A., GLEICHER M.: Scatterplots: Tasks, data, and designs. *IEEE Transactions on Visualization and Computer Graphics 24* (2018), 402–412. 2

[Sno55] SNOW J.: *On the mode of communication of cholera*. John Churchill, 1855. 2

[SPFCC15] STRANDBURG-PESHKIN A., FARINE D. R., COUZIN I. D., CROFOOT M. C.: Shared decision-making drives collective movement in wild baboons. *Science 348*, 6241 (2015), 1358–1361. 3

[TGC03] TRUTSCHL M., GRINSTEIN G., CVEK U.: Intelligently resolving point occlusion. In *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on* (2003), IEEE, pp. 131–136. 2

[TGK*14] TOMINSKI C., GLADISCH S., KISTER U., DACHSELT R., SCHUMANN H.: A survey on interactive lenses in visualization. *EuroVis State-of-the-Art Reports 3* (2014). 2

[Wil06] WILKINSON L.: *The grammar of graphics*. Springer Science & Business Media, 2006. 2