Binning Strategies and Related Loss for Binned Scatterplots of Large Data

Author A[1], Author B[2,3], Author C[2]
[1]Affiliation X
[2]Affiliation Y
[3]Affiliation Z

December 5, 2016

Abstract

Dealing with the data deluge of the Big Data Age is both exciting and challenging. The demands of large data require us to re-think strategies of visualizing data. Plots employing binning methods have been suggested in the past as viable alternative to standard plots based on raw data, as the resulting area plots tend to be less affected by increases in data. This comes with the price of the loss of information inherent to any binning scheme. In this paper we discuss binning algorithms used in the construction of binned scatterplots. We define functions to quantify the loss of spatial and frequency information and discuss the effects of binning specification on loss in the framework of simulation and case studies. From this we provide several practical suggestions for binning strategies that lead to binned scatterplots with desirable visual properties.

Keywords: binned scatterplots, visual loss, aggregation, graphics

1 Introduction

Technological advances have facilitated collection and dissemination of large data as records are digitized and our lives are increasingly lived online. According to an EMC report in 2014 "the digital universe is doubling in size every two years and will multiply 10-fold between 2013 and 2020 - from 4.4 trillion gigabytes to 44 trillion gigabytes". [1] This "Data Deluge" of the Big Data Age (NY Times, Feb 2012) poses exciting challenges to data scientists everywhere: "It's a revolution ... The march of quantification, made possible by enormous new sources of data, will sweep through academia, business and government. There is no area that is going to be untouched"– Gary King, Harvard Institute.

Data sets with millions of records and thousands of variables are not uncommon. Friedman (1997) proposed in his paper on data mining and statistics that "Every time the amount of data increases by a factor of ten, we should totally rethink how we analyze it". Jacobs (2009) echoed the sentiment, stating that "big data should be defined at any point in time as *data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time*". The same holds true for visualizations. With a 100-1000 fold increase in the amount of data, the utility of some of our most commonly used graphical tools, such as scatterplots, deteriorates quickly (Unwin et al., 2006a).

Area plots, such as histograms, do not tend to be as affected by increases in the amount of data because they display aggregations instead of raw data. By using binning strategies and the principles for displaying information in area plots, scatterplots can again become useful instruments for large data settings (Unwin et al., 2006a).

In this paper we describe first the inadequacy of traditional scatterplots in large-data situations. We discuss different binning algorithms use in the construction of binned scatterplots and the *loss of information* inherent to binning. We will then explore the effects of binning specification on the properties of binned scatterplots through simulation and real-data case studies. We conclude with several practical suggestions for binning specifications for creating binned scatterplots that have desirable visual properties.

[1] Access 12/18/2015 at `http://www.emc.com/about/news/press/2014/20140409-01.htm`

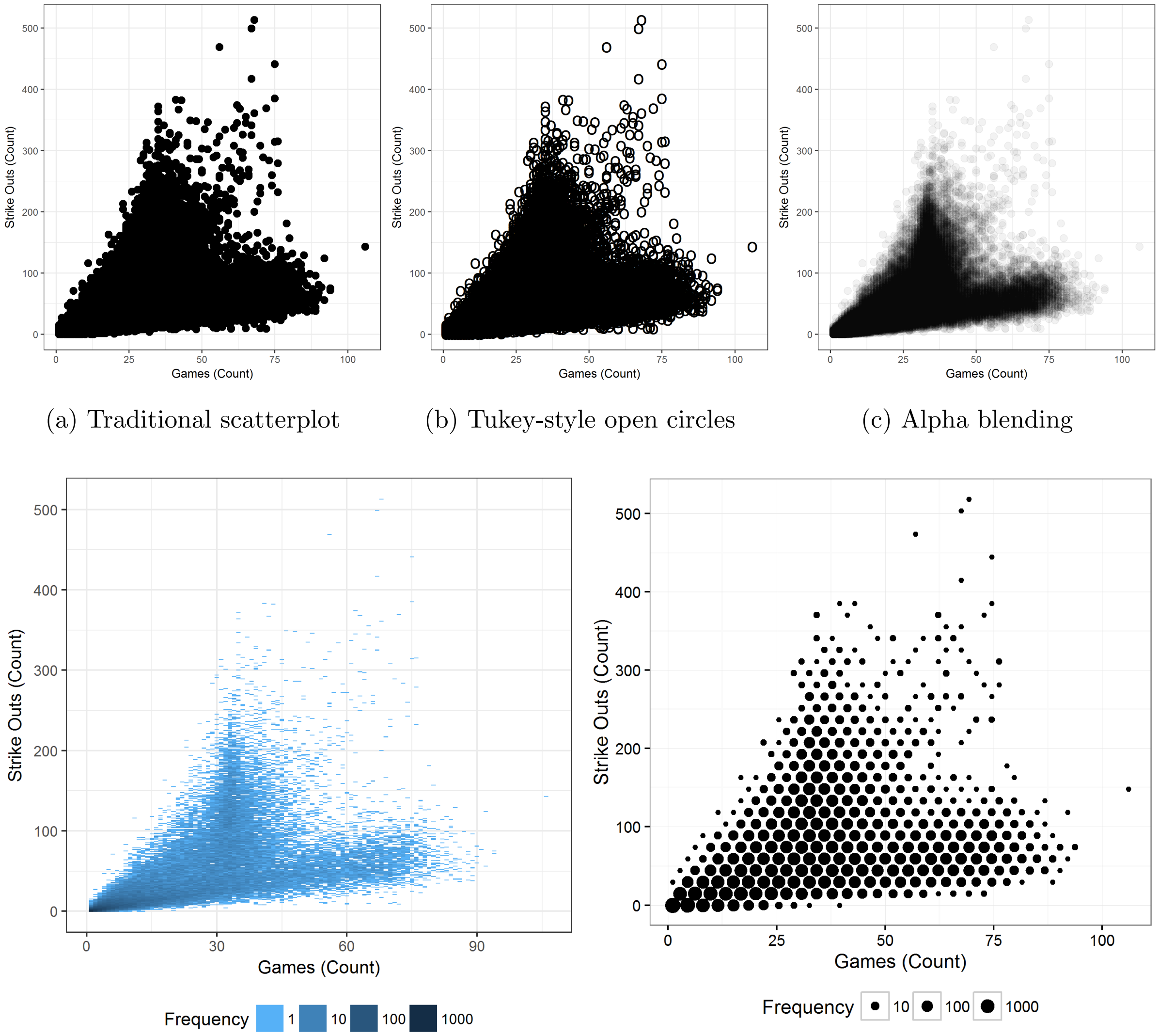
2 Scatterplots for Large Data Sets

In the case of modestly sized data, scatterplots are great tools for showing bivariate data relationships. With large data, scatterplots suffer from over-plotting of points, which masks relevant structure. Figure 1 shows an example taken from baseball statistics. The scatterplot shows 144 seasons (from the years of 1871 – 2014) of pitching statistics for every baseball pitcher as published in Sean Lahman's Baseball database.[2] The number of games played in a season is plotted against the number of strikeouts a pitcher threw over the course of a season. While the data set is only medium sized with 42583 observations, it already shows some of the break-down patterns scatterplots experience with large data.

The top row of Figure 1 demonstrates several variants of the traditional scatterplot where each observation is plotted using distinct points. The traditional, solid point, scatterplot in Figure 1(a) shows a triangular structure is apparent with some outliers at a medium number of games and high number of strikeouts; however the density within the triangular mass of points is indistinguishable. Tukey (1977) suggested the use of open circles (see Figure 1(b)) to mitigate the problem of over-plotting. A modern alternative to open circles is alpha blending (see Figure 1(c)) which renders points as semi-transparent to provides more visibility of underlying points. The data set is large enough that neither alpha blending nor open circles are completely effective, and so we must pursue a different strategy which can provide better information about the relative density of points at a given location.

Other scatterplot adaptations have been introduced that avoid over-plotting by manipulating the display of the points by distorting the locations or the scales. Generalized scatterplots (Keim et al., 2010) display all individual observations, including those sharing identical coordinates, and use distortion of the point locations by having points repel one another to avoid overlapping. An extension of generalized scatterplots uses clustering and local principal components to allow ellipsoid oriented distortion to display local correlation structure in the data (Janetzko et al., 2013). Variable-binned scatterplots (Hao et al., 2010) break the display into a non-uniform rectangular grid and re-size the rows of cells according to density of points. This variable binning fragments the continuity of the axes into segments

[2] Data Accessed 12/16/2015 at `http://www.seanlahman.com/baseball-archive/`



(a) Traditional scatterplot

(b) Tukey-style open circles

(c) Alpha blending

(d) Rectangular minimally binned scatterplot

(e) Hexagonal binned bubble plot

Figure 1: Traditional and adapted scatterplots for games vs. strikeouts data. Note improved density information in the aggregation-based plots.

on different scales and also does not deal with points at identical coordinates. Generalized and variable-binned scatterplots make fine data structure more visible and allow color to be reserved for a third variable instead of frequency; however, the distortion of the point locations and/or axes warp the visual display of the association between the two primary variables.

Another approach is to reduce the graphical complexity by plotting binned aggregations of the data, namely frequencies, as opposed to plotting every observation as an individual point. This has the additional advantage of reducing the size of the stored data necessary for the construction of the plot, as only the bin centers and the bin frequencies must be stored. Wickham argues for a "bin-summarize-smooth" procedure to be applied to the visualization of big data and he notes that simple summary functions, such as counts, scale well with the size of data (Wickham, 2013). Liu, Jiang and Heer employ the computational benefits of binning for their interactive big data visualization program, *imMens* (Liu et al., 2013).

Histograms are a simple example of a plot that can be built using binned aggregations of the data; in their case the bin locations and bin counts act as a set of sufficient statistics necessary to reconstruct the plot. A natural extensions of histograms to higher dimensions is to form a tessellated grid on a two dimensional Cartesian plane using some other attribute, such as color or size or 3D renderings to provide joint density information within each grid cell, known as a tile. A *sunflower plot* uses symbols that increase in complexity proportional to the number of points in each bin. Due to difficulty in rendering discernible shapes in limited space, sunflower plots are only useful when the number of points in each bin remains reasonably small. A *bubble plot* is a binned data plot that scales the size of a filled circle in proportion to frequency. Bubble plots were first used by William Playfair (Playfair, 1786; Playfair et al., 2005). A *binned scatterplot* uses shading to provide frequency information, with tiles (rather than bars in a histogram) at the bin center, similar to a two-dimensional histogram viewed from above.

Figure 1 contains examples of a rectangular binned scatterplot with frequency encoded as color (d) and a hexagonal binned bubble plot with frequency encoded as point size (e); both of which are more effective at displaying the shape of the joint density and preserving outliers than any of the scatterplots shown in Figure 1(a-c). Bubble plots is prone to suffer from the Hermann-grid illusion (Hermann, 1870), where the white spaces between circles on the evenly spaced grid appear shaded due to an optical illusion. The rectangular bins in Figure 1(d) are one game by one strikeout in size which matches the resolution of the recorded data; this is referred to as a *minimally binned scatterplot.* With only a single unique coordinate pair exists within each bin the tiles are comparable to points in a traditional scatterplot; however, the unique coordinate pairs in a traditional scatterplot are shaded in a binary manner with no indication of overlapping observations. Alpha blending as used in Figure 1(c) is akin to bin shading; however the frequency mapping is imperfect because as soon as overlapping points surpass full opacity the perceivable frequency information is truncated. By explicitly shading bins according to frequency, more information is preserved than in a traditional scatter plot, as the frequency domain provides visual weight to tiles which may represent more points.

The inner structure of the baseball data is only apparent in the binned scatterplot and the bubble plot. The joint density consists two distinct ridges following two lines with very different slopes. A low ridge with high games and lower strikeout rates, and a high ridge with fewer games played but high strikeout rates. Closer investigation of additional variables reveals that this split in the density ridges corresponds mainly to pitchers in the pre-modern and modern era's of baseball; players in past had much shorter seasons (in 1876 only 70 games were played in a season, as opposed to 162 in 2009), and pitchers are disadvantaged in the modern era due to substantial qualitative improvements in bats.

For extremely large data sets, binned scatterplots are a more useful visualization of two-dimensional density information than the scatterplot, and are less computationally demanding, as not every single point in the data set has to be rendered separately. In order to explore the properties of binned scatterplots we must specify binning algorithms by which to aggregate.

3 Binning Algorithms

Binning algorithms used in making distributional approximations can be traced back to Pearson's work with the binomial approximation to the normal, where he mentions the need to define an origin and binwidth for segmenting the normal distribution (Pearson, 1895). Sturges followed with an early work in formalizing histograms specification (Sturges, 1926). More recently Scott has presented discussion on the importance of binning specification in the creation of histograms to appropriately display one dimensional density approximations (Scott, 1979). Scott (1992) extends to the properties of multivariate binning strategies.

Binning in dimensions X and Y provides us with a more condensed form of the data that ideally preserves both the joint distribution as well as the margins, while reducing the amount of information to a fraction of the original. Binning is a two-step procedure: we first assign each observation (x, y) to a bin center (x^*, y^*), and in a second step we count the number of observations assigned to each unique bin center; resulting in reduced data triples of the form (x^*, y^*, c), where c is the number of all observations assigned to bin center (x^*, y^*).

We will proceed with rectangular bins for simplicity, but other binning schemes, such as hexagonal bins (Carr et al., 1987) are also common. While hexagonal binning has been shown to have slightly better graphical properties (Scott, 1992); rectangular bins are advantageous because bins in x and y are orthogonal to each other, thus we can present the one-dimensional case which will easily generalize to two or more dimensions (Unwin et al., 2006b). We will however only consider binning in up to two dimensions, X and Y. The algorithms we discuss are immediately applicable to higher dimension, but we do not feel that the paper would benefit from a more general discussion.

For the univariate case with observations, x_i for $i \in \{1, \ldots, n\}$, binning algorithms require a set of bin centers x_j^* for $j \in \{1, \ldots, J\}$ and a binning function $b_X(.) : x_i \rightarrow x_j^*$ that maps observations to bin centers. What we will refer to as *general rectangular binning* accomplishes this by defining a sequence of J adjacent intervals, $(\beta_{j-1}, \beta_j]$ for $j \in \{1, \ldots, J\}$, which span over the range of the data. Note that half open intervals are used such that any observation falling on a bin boundary is assigned to a unique interval. Values x_i exactly equal to the lowest bin boundary β_0 are grouped into the first bin to close the leftmost bound. Each observation is then mapped to a bin center, x^*_j; the midpoint for the interval to which the observation belongs.

This is expressed mathematically using the binning function $b_X(.) : x_i \to x^*_j$ defined as

$$b_X(x_i) = \begin{cases} x^*_1 & \text{for all } x_i = \beta_0 \\ x^*_j & \text{for all } x_i \in (\beta_{j-1}, \beta_j] \end{cases} \tag{1}$$

Standard rectangular binning is a special cases of general rectangular binning that uses intervals of equal size for all bins; thus only the origin of the first bin, β_0, and binwidth, ω_X, need to be specified. Standard rectangular binning is necessarily used in the construction of histograms (Pearson, 1895); the consistent binwidth makes the display of frequency proportional to density . Fixed width binning procedures are also highly computationally efficient (Wickham, 2013).

As an alternative to the rectangular binning process, we propose a *random binning* algorithm which utilizes a non-deterministic bin function $b^r_X(\cdot)$ to randomly assigns an observation, x_i, to a bin center, x^*, from a set of possible bins. In this paper, we will consider the simplest case of just two bins, so that without loss of generality we can assume that x_i lies between bin centers x^*_j and x^*_{j+1}. The bin function assigns x_i to a bin center with a probability inversely proportional to the distance to that bin center; the closer a value is to a bin center, the higher the probability the value is assigned to that bin center. More formally,

$$b^r_X(x_i) = \begin{cases} x^*_j & \text{with probability } (x^*_{j+1} - x_i)/(x^*_{j+1} - x^*_j) \\ x^*_{j+1} & \text{with probability } (x_i - x^*_j)/(x^*_{j+1} - x^*_j) \end{cases} \tag{2}$$

for $x_i \in [x^*_{j+1}, x^*_j]$. In Table 1 we note that this random binning algorithm does not specify bin boundaries; only a sequence of bin centers. This method is easily extensible to also map x_i into more than two bins and can accommodate non-uniform distribution of bin centers. The impetus for developing the random binning assignment was to soften the hard breaks associated with one-sided interval assignment for scenarios where many points fall directly on bin boundaries by allowing points to be divided stochastically. The bin boundaries and centers for standard and random rectangular binning algorithms can be found in Table 1.

	Bin Boundaries	Bin Centers
General	$\{\beta_j \mid \beta_j > \beta_{j-1}\}$	$\{x_j^* \mid x_j^* = (\beta_{j-1} + \beta_j)/2\}$
Standard	$\{\beta_j \mid \beta_j = \beta_{j-1} + \omega_X\}$	$\{x_j^* \mid x_j^* = \beta_{j-1} + \omega_X/2\}$
Random	—	$\{x_j^* \mid x_j^* > x_{j-1}^*\}$

Table 1: Rectangular and Random Binning Specifications

Quantile binning is another option that divides the range of the observations into bins each containing an equal number of points. The j^{th} bin interval takes the form $(Q_X((j-1)/J), Q_X((j)/J)]$, where $Q_X(p)$ is the the p^{th} empirical quantile using the inverse empirical distribution function (Hyndman and Fan, 1996). Note that this binning approach is *not* desirable for spatially visualizing density patterns, as it effectively balances the frequency counts in all bins; it does however have desirable properties for binned scatterplots that employ a second stage of binning to create discrete shade scheme for displaying grouped bin frequencies, which will be discussed in Section 4.2.

3.1 Extension to Two Dimensional Binning

The standard and random binning algorithms are easily extended to higher dimensions when binned orthogonally (Unwin et al., 2006b). For the purposes of creating binned scatterplots we will specify extension to rectangular binning in two dimensions. In this case we wish to assign data pairs (x_i, y_i) to bin centers of the form (x_j^*, y_k^*), with $j \in \{1, \ldots, J\}$ and $k \in \{1, \ldots, K\}$,where J and K are the number of bins in the X and Y dimensions, respectively. The (j,k) pairs that index the bin centers can be linearized to a single index such that $\ell = j + J(k-1)$; thus making j the fast running index and k the slow running index. With this linearized index for all bins we now have a set of bin centers of the form (x_ℓ^*, y_ℓ^*), with $\ell \in \{1, \ldots, \mathscr{L}\}$, where $\mathscr{L} = J \cdot K$.

The standard rectangular binning function $b(.) : (x_i, y_i) \to (x_\ell^*, y_\ell^*)$ is defined as

$$b(x_i, y_i) = (b_X(x_i), b_Y(y_i)) \tag{3}$$

where $b_X(x_i)$ and $b_Y(y_i)$ are the univariate standard binning algorithms for the X and Y dimensions respectively. The random rectangular binning function, $b^r(\cdot) : (x_i, y_i) \rightarrow (x^*_\ell, y^*_\ell)$ is similarly defined as

$$b^r(x_i, y_i) = (b^r_X(x_i), b^r_Y(y_i)) \tag{4}$$

where $b^r_X(x_i)$ and $b^r_Y(y_i)$ are univariate random binning algorithms for each dimension. Figure 2 provides an illustration of each binning process extended to a two dimensional situation.

3.2 Binned Data Reduction

The second stage of binning requires a frequency breakdown of the number of observations associated with each bin center, forming reduced data triples, (x^*, y^*, c), where c is the number of all observations assigned to bin center (x^*, y^*). Table 2 makes use of a small set of simulated data to show the progression from the original data (a), to the binned data (b), to the reduced binned data (c). The reduced binned data is sufficient for constructing the binned scatterplot. In cases of large data, binning greatly reduces the storage size for the information and the computation time needed to construct a binned scatterplot. Note that numerical attributes other than frequency of the binned data may also be recorded during binning, however only frequency is required to construct a binned scatterplot. Data reduction comes at the expense of spatial information of any of the individual points, which can only be recovered when minimally binned at the resolution of the data. The loss of information incurred from binning will be explored in following sections.

4 Loss due to Binning

Problems with large data in scatterplots arise from over-plotting, which is a form of implicit data aggregation. In order to keep track of the number of observations near a given location, we switch to a weighted visual display which explicitly aggregates the data. The reduced binned data carries the sufficient information necessary to render the binned scatterplot. Making the data aggregation explicit allows us to calculate the loss we experience.

Loss of information occurs during the binning and rendering process. For the remainder

x	y
-7.7325	-9.6340
-8.1176	-1.4529
-5.8996	-3.2033
-7.0375	-5.5563
-3.6354	-3.9315
-8.7639	0.9874
-2.9781	8.6802
0.8210	-8.6118
5.4477	-8.4555
4.6849	-5.6620
9.4785	1.1133
1.7579	5.3759

(a) Original Data

$b_X(x)$	$b_Y(y)$
-5	-5
-5	-5
-5	-5
-5	-5
-5	-5
-5	5
-5	5
5	-5
5	-5
5	-5
5	5
5	5

(b) Binned Data Centers

x^*	y^*	c
-5	-5	5
-5	5	2
5	-5	3
5	5	2

(c) Reduced Binned Data

Table 2: Original, binned and reduced binned data tables, using standard rectangular binning with origin $(\beta_{0,x}, \beta_{0,y})$ = (-10,-10) and binwidths $\omega_x = \omega_y = 10$.

of the paper we will assume that we are using shade in binned scatterplots to represent frequencies. We distinguish two sources of loss in the construction of a binned scatterplot:

- *Spatial Loss*, L^{S}, occurs when points (x_i, y_i) for observations $i \in \{1, \ldots, n\}$ in the data set are reduced to a set of tiles centered at (x^*_ℓ, y^*_ℓ) for bins $\ell \in \{1, \ldots, \mathscr{L}\}$. By displaying frequency information using shaded tiles instead of individual points there is a loss of information about the exact location of the points.

- *Frequency Loss*, L^{F}, occurs when bin counts, $c_\ell \in \{1, \ldots, \mathscr{L}\}$ are not mapped to a continuous shading scale. While shade can be *rendered* continuously in HSV color space, thus representing frequency exactly, a human reader can not *extract* this information at the same precision due to limitations of human cognition. In order to model these limitations we introduce a second stage of binning by using a discrete color scale for displaying binned frequencies, $b_C(c_\ell)$, $\ell \in \{1, \ldots, \mathscr{L}\}$.

Note that while the losses from creating a binned scatterplot may turn out to be substantial, they present a huge gain with respect to an traditional scatterplot, where density information is implicitly masked in large data situations. The idea of loss from one-dimensional binning was explored by Scott using mean integrated squared error as the loss function to be optimized by the choice of the number of bins in the construction of histograms (Scott, 1979). He later extended this discussion to two-dimensional binning, where he compared the mean integrated squared error (MISE) loss for hexagonal, rectangular and triangular binning; finding that hexagonal and rectangular binning performed similarly, both far superior to triangular binning (Scott, 1992). Scott's MISE assesses the difference between the empirical density and the binned density approximation, via integration of the squared deviation over $\mathbb{R}^2$. Whereas, we quantify binning losses using the spatial displacement of the points to bin centers using euclidean distances, which better reflects the visual information lost in the shift from the traditional to binned scatterplots.

4.1 Spatial Loss

When the individual points of a scatterplot are collapsed to bin centers to be displayed as tiles in a binned scatterplot there is a loss of the location information. The tile size impacts the precision with which a user can read out location related attributes of the scatterplot; such as fine density patterns or outlier identification. This loss can be expressed as the Euclidean distance between points and the visual center of the tiles (i.e. the bin centers). Note that other distance metrics could be used, but the Euclidean distance has a desirable interpretability in $\mathbb{R}^2$. The *total spatial loss*, L^S, is defined as

$$L^S = \sum_{i=1}^{n} L_i^S = \sum_{i=1}^{n} \sqrt{\left(x_i - b_X(x_i)\right)^2 + \left(y_i - b_Y(y_i)\right)^2} \qquad (5)$$

where L_i^S is the loss in the ith observation. Figure 2 visually displays the spatial loss for the data from Table 2 as a result of standard rectangular binning. Observations (x_i, y_i) and bin centers (x_ℓ^*, y_ℓ^*) are displayed as black points and gray crosses, respectively. The length of line segments connecting these represent L_i^S, the spatial loss for each observation; thus, the combined length of all line segments represents the total spatial loss, L^S. For random assignment the total spatial loss can be calculated by simply replacing the standard binning function with the random binning function in Equation 5.

The total spatial loss for randomly binned data is visualized in Figure 2. We see that in random binning there are pairs of points – like those with dashed line segments – that have

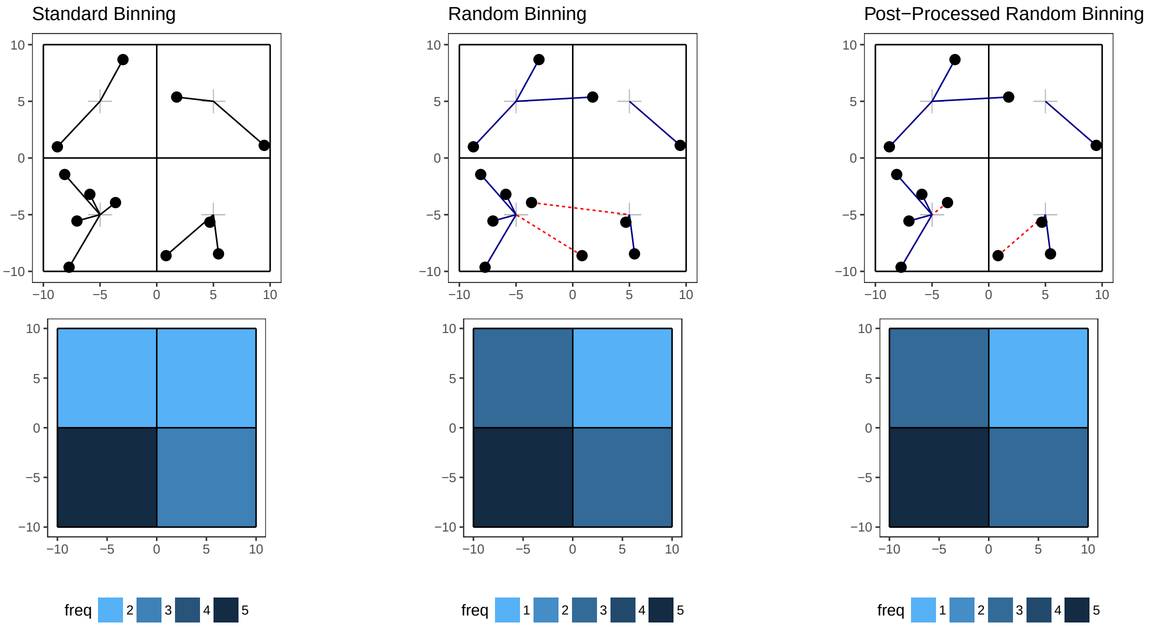


Figure 2: Visualization of spatial loss for same data using standard, random and post-processed random binning algorithms.

been allocated in such a way that they are closer to their partner's bin center than their own. We see that the binned scatterplot remains identical if these points are swapped back to the closer centers; however, the total spatial loss is smaller after the random allocation is post-processed. To appropriately reflect the perceived loss of spatial information we use the *net spatial loss*, L^S_{net}, which is the minimum total spatial loss from all binning allocations that result in the same reduced binned data, and thus the same binned scatterplot. For standard binned data, the net spatial loss is always equivalent to the total spatial loss due to the deterministic bin allocations. For random binned data, the net spatial loss is achieved by first exchanging bin assignments for all pairs of points in neighboring bins that exist further from their own bin center than from their partner's bin center, then calculating the total spatial loss from this post-processed binned data.

The spatial loss is a Euclidean distance, but the units affiliated with this distance are based on the units on which each variable is recorded. If the two variables in the binned scatterplot share the same units this leads to direct interpretability of the spatial loss. However, if the two variables do not share the same units or the same magnitude of values it is advisable to standardize the variables prior to binning, thus making the spatial loss more universally interpretable as a distance in units of standard deviations.

4.2 Frequency Loss

Bin counts can be displayed using a continuous shading scale in HSV color space (Healey and Enns, 1999) and thus we can theoretically map frequency to shade perfectly. While the tiles of the binned scatterplot would be *rendered* precisely, the ability of a human with average vision to extract that information by visually mapping the tile shade to a frequency scale in the plot legend is largely imprecise. Color perception of shade is influenced by the contrasting effect of surrounding shades, allowing an inaccurate mapping of tile shade to the corresponding shade in the plot legend (Bartleson and Breneman, 1967; Adelson, 1993; Fairchild, 2013). See Figure 3 for a demonstration of context sensitivity of colors. It is therefore not realistic, to expect readers to be able to decode frequency or accurately compare bin frequencies from a continuous shade scheme, even though theoretically we can perceive shades continuously (Leong, 2006).

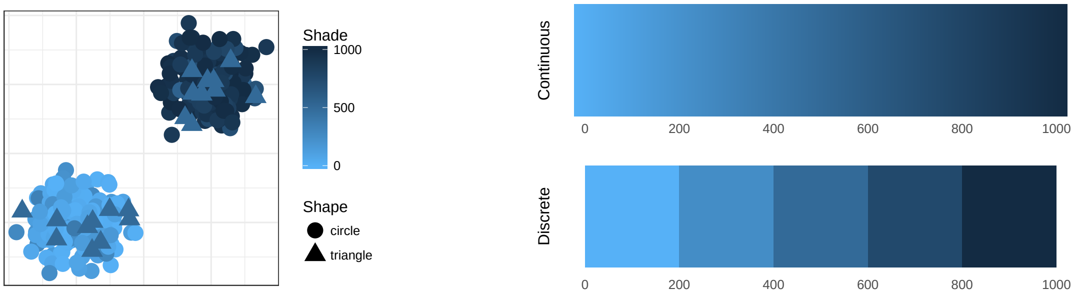


Figure 3: (Left) Q: Which triangles are darker? A: All triangles same shade = 500. (Right) Continuous and standard rectangular binned frequency color scales.

Whenever the shading scheme for rendering the counts in each bin is discretized there is a loss of frequency information. We can model this as a second stage of binning; wherein the bin counts, c_ℓ, for bins $\ell \in \{1, \ldots, \mathscr{L}\}$ are placed into frequency bins using any of the previously discussed univariate binning algorithms in Section 3. Figure 3 provides a visual example of a discrete color palette resulting from frequency binning.

Research suggests that even under optimal conditions, we can effectively compare only about seven color hues simultaneously, and that we are even more limited in terms of distinguishing shade (Healey and Enns, 1999). This provides a physical upper limit on the amount of frequency variation we can perceive through color. As a result, a frequency binning which produces seven of fewer frequency categories is preferable.

The goal of binning the frequencies and using a ordinal shading scheme is to quantify the imprecision in visually extracting frequency information. It does so by using shade to display binned frequencies, $b_c(c_\ell)$ instead of the true frequencies, c_ℓ. The *total frequency loss*, L^F, is defined as

$$L^F = \sum_{\ell=1}^{\mathscr{L}} L_\ell^F = \sum_{\ell=1}^{\mathscr{L}} (c_\ell - b_c(c_\ell))^2 \tag{6}$$

where L_ℓ^F is the frequency loss for the ℓ^{th} bin. Note that this is effectively a sum of squared deviations between true frequencies to the centers of frequency bins. For example, in the discretized color scale in Figure 3 that uses bins with bounds $\{0, 200, 400, 600, 800, 1000\}$ the frequency deviation for each spatial bin would be compared to the corresponding bin centers at $\{100, 300, 500, 700, 900\}$. While this numerical assessment of frequency loss does not exactly account for limitations in human perceptual ability, it does provide a more realistic model for the loss in perception that does occur.

Frequency data consists of counts, which commonly exhibit skew densities, i.e. there are usually a lot of bins with small bin counts and a few bins with extremely large frequencies. The use of quantile binning is promising in the case of frequency binning because it seeks to place the same number of bins in each shaded group. An alternative is to use a log transformation which produces a more symmetric distribution of frequency information, increasing perceptual resolution. This is consistent with the Weber-Fechner law which suggests that increased stimulus intensity is perceptually mapped on the log scale (Goldstein, 2007). Using a logarithmic mapping of frequency to the shade aesthetic provides a more natural perceptual experience and simultaneously increases the perceptual resolution of the graph. The *log frequency loss*, $L^{\log F}$, is defined as

$$L^{\log F} = \sum_{\ell \in \mathscr{L}^*} L_\ell^{\log F} = \sum_{\ell \in \mathscr{L}^*} (\log(c_\ell) - b_c(\log(c_\ell)))^2 \tag{7}$$

where $\mathscr{L}^*$ is the index set for all non-empty bins, which is done to avoid asymptotic problems from log transforming bin counts of zero.

5 Exploring Properties of Loss

Binning data for the purpose of creating a binned scatterplot requires a choice of algorithm as well as a choice of parameters associated with that binning algorithm. This section aims to compare binning algorithms and identify the best parameter choices for minimizing loss under a number of distributional scenarios. Some choices may be proven optimal through analytical properties, while other are data dependent and require empirical exploration of loss from binning. Whether analytical or empirical, data is needed to demonstrate how loss is impacted by binning choices.

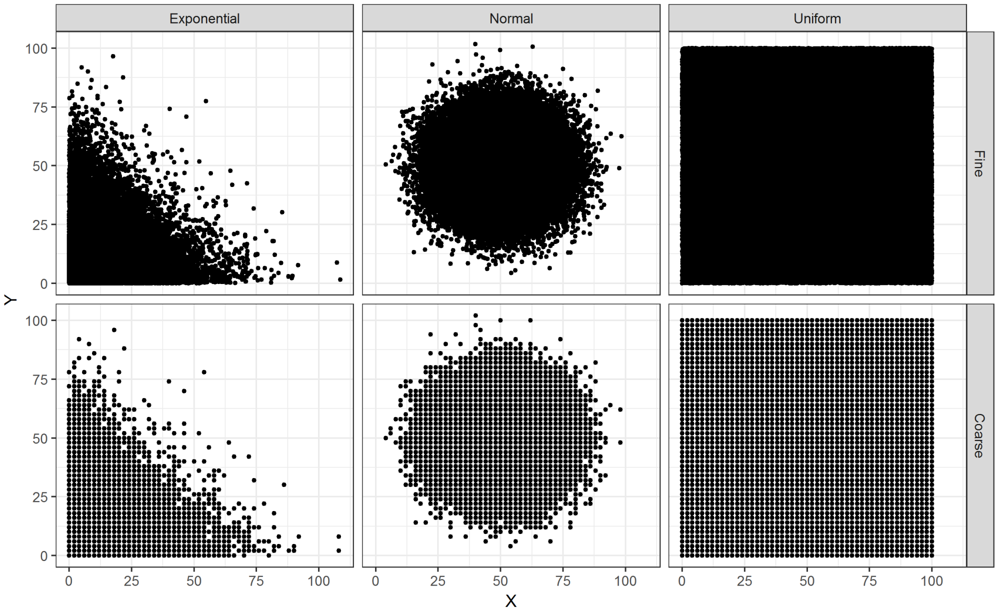


Figure 4: Scatterplots of fine and coarse versions of the simulated bivariate data.

Data sets were simulated from bivariate distributions to be used throughout this section: Exponential, Normal and Uniform. 100,000 observation pairs were simulated for each data set from the following distributions:

- Set I: $x_i \sim$ iid Exp$(\lambda), y_i \sim$ iid Exp(λ) with $\lambda = 11$
- Set II: $x_i \sim$ iid Normal$(\mu, \sigma^2), y_i \sim$ iid Normal(μ, σ^2) with $\mu = 50$ and $\sigma = 11$
- Set III: $x_i \sim$ iid Uniform$(a, b), y_i \sim$ iid Uniform(a, b) with $a = 0$ and $b = 100$

The parameters were selected to have data values roughly span the region $[0, 100]^2$. The simulated data can be found in the top row of Figure 4. These simulated data sets are from continuous distributions, and thus the values are recorded to many decimal places; 6 decimal places in our simulate data. Real data is recorded to only the number of digits that measurement precision allows, and in many cases rounded even further.

The *data resolution* is defined as the smallest increment between successive data values. To observe loss from binning under more realistic conditions, we create three data sets by rounding the values from the originally simulated data sets to the nearest even number, thus a data resolution of 2 units in each dimension. This coarse version of the original simulated data is displayed in the bottom row of Figure 4. By exploring the loss properties for both the *fine* and *coarse* versions of the data, we identify which binning options are robust to the resolution at which data is recorded.

Dist-Res	Outlie	Skew	Clump	Sparse	Striate	Convex	Skinny	String	Mono.
Exp-Fine	0.15	0.67	0.01	0.02	0.02	0.47	0.55	0.36	0.00
Exp-Coarse	0.15	0.71	0.02	0.02	0.03	0.51	0.52	0.27	0.00
Norm-Fine	0.07	0.55	0.00	0.02	0.01	0.60	0.31	0.39	0.00
Norm-Coarse	0.08	0.59	0.01	0.02	0.02	0.60	0.33	0.35	0.00
Unif-Fine	0.00	0.52	0.00	0.03	0.00	0.68	0.15	0.36	0.00
Unif-Coarse	0.00	0.56	0.00	0.03	0.01	0.68	0.14	0.38	0.00

Table 3: Scagnostics scores for the six simulated datasets used in empirical exploration of spatial loss in binned scatterplots.

These bivariate distributions were selected for the variety in their defining characteristics. *Scagnostics* - a portmanteau for scatterplot diagnostics - are formalized metrics that have been used to numerically summarize key attributes of scatterplots, including: *outying,* *skewness, clumpiness, sparsity, striation, convexity, skinniness, stringiness* and *monotonicity* (Wilkinson et al. 2005;Wilkinson and Wills 2008). The scagnostic scores in Table 3 were calculated for the simulated data sets using the `scagnostics` package in R (Wilkinson et al., 2015). They show that these examples explore a variety of *skewnesses*, *convexity* and *skinniness*. This exploration uses bivariately independent distributions, thus the monotonicity and clumpiness are both near zero for these examples. The coarse data is rounded into striatations by definition, however the striation scagnostic score is low due to the relatively small scale of striation versus the magnitude of variability in each distribution. These minor striations will still prove problematic when binning is inconsistent with the data resolution.

5.1 Rectangular Binning Specifications and Spatial Loss

For rectangular binning specification options include the type of algorithm, location of the origin and binwidths for each dimension. To explore the spatial loss properties under different binning approaches we begin with a comparison of standard and random binning. Figure 5 displays the net spatial loss from binning the fine resolution simulated data with standard and random binning algorithms using square bins with a sizes ranging from 2 units^2 to 20 units^2. For equal bin sizes, the net spatial loss under standard binning is always less than or equal to the net spatial loss under random binning. This is because the minimal spatial loss for each data point under random binning is to allocate to the nearest bin center, which is how the point would be allocated in standard binning. The net spatial loss from random binning becomes increasingly costly as bin sizes increase, as indicated by the widening gap between the lines and the steeper slopes.

If we view the binned scatterplot as a visual estimator of the bivariate density then we may consider a few desirable properties of estimators: unbiasedness, consistency and efficiency. Binned scatterplots shift visual emphasis from true location of individual points to the geometric centers of tiles, making them visually biased displays of density. However, as bin sizes become increasingly fine, the bias decreases; with no visual bias when minimal binning perfectly matches bins with the data resolution the density estimate. Also, the density estimation more perfectly reflect the bivariate density as the sample size increases;

thus making the binned scatterplot a consistent visual estimator. We may also consider a minimal binned scatterplot as a spatially efficient estimator because it minimizes spatial loss in the visual density estimator. It is worth noting that a density estimate requires the combination of spatial and frequency information and that the estimation properties were considered only through the lens of spatial loss. This makes the assumption that frequency information is rendered through a precise continuous mapping of bin frequencies.

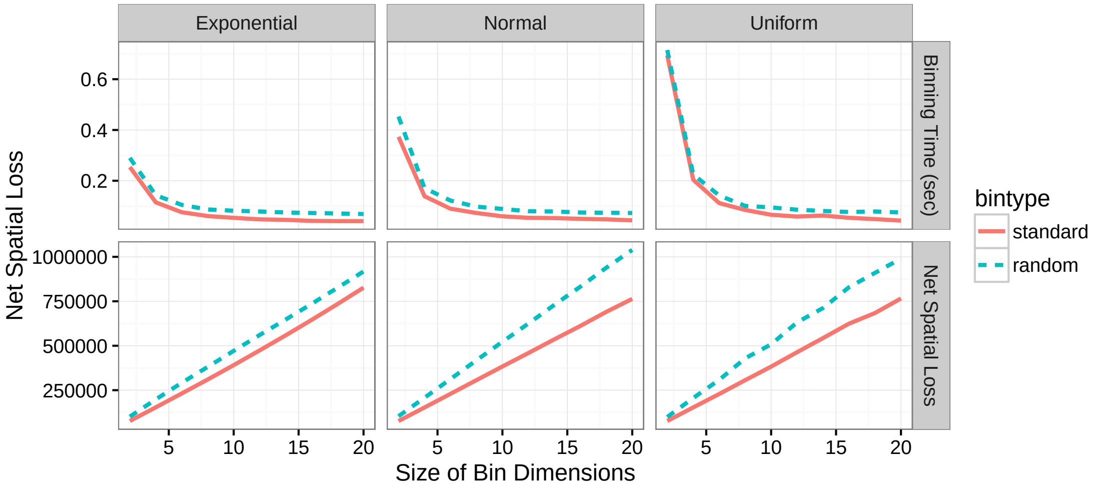


Figure 5: Lineplots for net spatial loss and computation times over a range of bin sizes for standard and random binning.

While using smaller bin widths leads to smaller spatial losses, reducing bin sizes comes at the cost of computation time; a potentially non-negligible consideration in settings with truly massive data. Figure 5 also shows that the computation time needed to bin the simulated data sets is a decreasing function of the bin size and that random binning is marginally slower than standard binning across all bin sizes.[3] Larger bins may also be reasonable to use if we are primary interested visualizing the large scale density structure and wish to smooth over fine structural noise.

While bin widths can be chosen as any positive real value, bin widths should be specified as an integer multiple of the resolution of the data because non-integer multiples of the data resolution lead to systematically different numbers of possible data values per bin. For bin

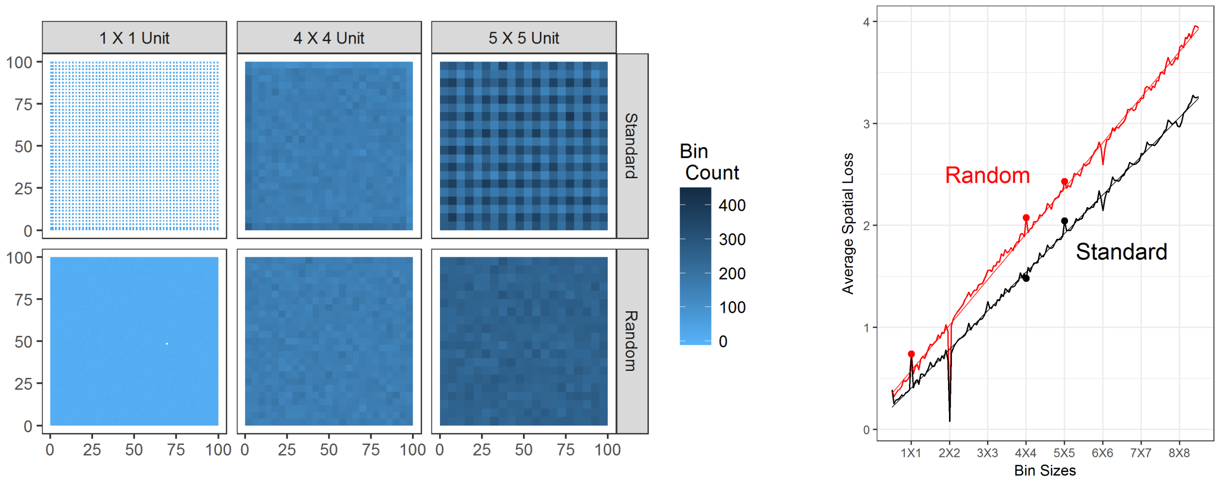
[3] Computation times using a commercial Asus laptop with an Intel Core i7 processor running at 2.80 GHz.

dimensions are smaller than the resolution of the data, there will be empty rows or columns of tiles in the binned scatterplot. While these gaps do exist in the data, they are undesirable because they create visual discontinuity that interferes with the interpretation of bivariate density and could also create the Hermann-grid optical illusion (Hermann, 1870; Spillmann, 1994). More seriously, bin dimensions that are non-integer multiples larger than the data resolution lead to *artificial striping* – an oscillating density pattern imposed by the binning that does not exist in the raw data.

To demonstrate the importance of properly selecting binwidths we consider the coarse version of the simulated bivariate-uniform data which is recorded to a resolution of two units in each dimension – thus even integer sized bins are most desirable. Figure 6(a) displays the binned scatterplots under several scenarios. Under standard binning we see the white-space gaps with one-by-one unit bins, vertical and horizontal artificial stripes with five-by-five unit bins, and the appropriate view of an evenly spread density with four-by-four unit bins. Figure 6(b) demonstrates that with coarse data resolution there are departures from the linear relationship between spatial loss and bin sizes; with small spikes in spatial loss where the bin size is highly misaligned with the data resolution and slight drops in spatial loss when aligned perfectly. In this we can see that artificial striping is a symptom of poorly aligned bin specification.

Note that random binning is effective at smoothing out the artificial striping patterns when many point fell along bin boundaries – accomplishing the motivation for its development. However due to inflated spatial loss due to the stochastic bin assignment, standard binning with properly selected bin dimension is the better option.

The binning origin can also influence the spatial loss in the binned scatterplot. For data with fine resolution compared to the bin dimensions, the origin is only largely consequential for distributions with high density near a natural boundary – e.g. weights of small items bounded below at 0 – where the origin should align with the boundary so that the outermost bins do not cover empty density regions. As an example, Figure 7 displays the binned scatterplots of ten-by-ten unit bins for the fine exponential data where the binning origin at (0,0) incurs seven percent lower spatial loss than for the (-9,-9) origin due to the heavy



(a) Binned scatterplots

(b) Spatial loss vs bin sizes

Figure 6: Binned scatterplots and spatial losses for various sized bins with coarse uniform data. Note the slight dip in loss at even integers, when bins are aligned with data resolution.

overlap into negative regions.

For data with a coarse resolution the location of the origin is important because the origin controls the proximity of possible data values to bin centers. We define the *origin offset* for each dimension, as the tuple, $(o_x,\, o_y)$, by which we offset the bivariate bin origin from the data minima, $(x_{(1)},\, y_{(1)})$. Thus the origin offset indicates the number of units in each dimension to shift the binning origin below the origin naturally encouraged by the data, resulting in $(\beta_{0,x}, \beta_{0,y}) = (x_{(1)}, y_{(1)}) - (o_x, o_y)$. It can be shown analytically that an origin offset of $(\alpha_x/2, \alpha_y/2)$ units minimizes the net spatial loss in the situation with the following three properties: (i) data are recorded to a resolution of α_x units in the X dimension and α_y units in the Y dimension, (ii) points are symmetric distributed within rectangular bins, (iii) the bin dimensions are integer multiples of α_x and α_y, respectively (see proof in Appendix A).

In practice the $(\alpha_x/2, \alpha_y/2)$ origin offset is found to be a reasonable binning choice for lowering spatial loss for coarse resolution bivariately symmetric data using bin dimensions that are integer multiples of α_x and α_y. Figure 8 shows how the net spatial loss changes as the origin offset is shifted while using standard rectangular binning for the coarse simulated data sets. Note that for simplicity, changes to the origin offset are made equally in each dimension.

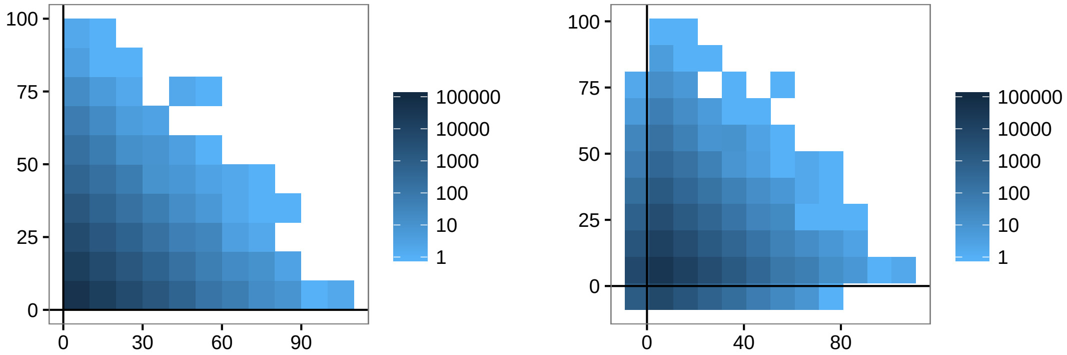


Figure 7: Binned scatterplots for the fine exponential data using standard binning with 10X10 square bins with origins at (0,0) and (-9,-9). The bold lines denote data lower bounds.

Since the coarse data has a 2X2 unit resolution, we pay special attention to an origin offset of (1,1) in order to assess how well the proposed default origin offset at ($\alpha_x/2$,$\alpha_y/2$) works in each scenario. The round glyphs indicate the origin offset where the net loss reaches an absolute minimum in the simulation. For the two symmetrically distributed data sets, normal and uniform, the origin offset of (1,1) was found to either minimize the net spatial loss or achieve a local minimum very near to the overall minimum (within a 0.2% increase from the minimum spatial loss) for each considered bin size. For the bivariately skewed exponential data, the origin offset of (1,1) minimized net loss for the smaller intervals but was not optimal for the largest intervals; 2.5% and 7% above the minimum spatial losses for the 8X8 and 10X10 unit bins, respectively.

5.2 Frequency Binning Specifications and Frequency Loss

The reduced binned data from spatial binning contains the center and count information for all bins. The bin frequencies may be mapped continuously to a precisely rendered shade, however it is naive to believe that human perception will be able to perfectly extract that information as a numeric value. There is implicitly loss of frequency information occurring when the shade of a tile is visually mapped back to a frequency through the use of a shading scale index. Bin frequencies may themselves be binned in order to discretize the color scale

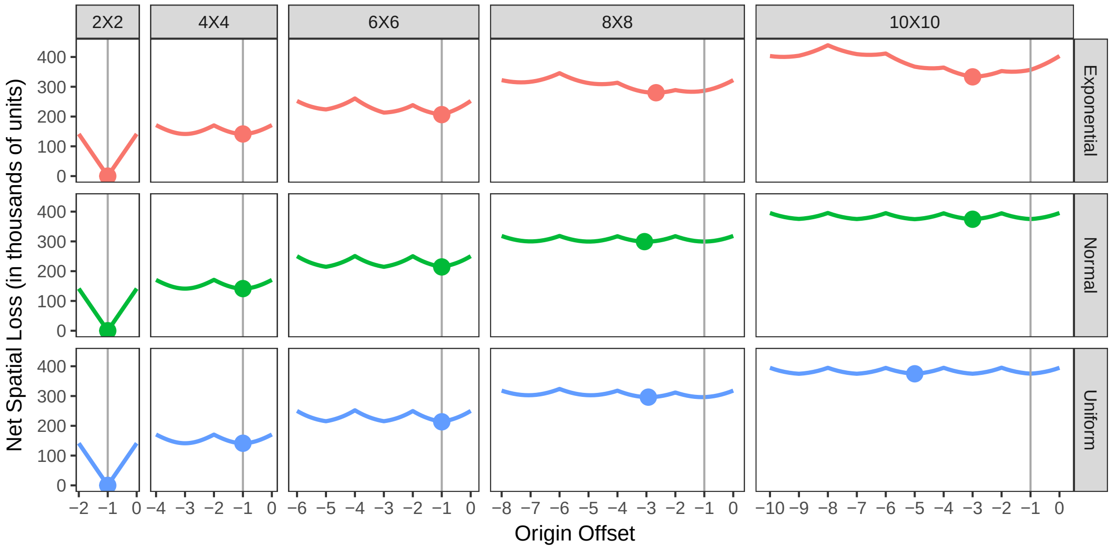


Figure 8: Net spatial loss 2X2 unit resolution data using various sized square bins over the range of possible origin offsets. The vertical gray lines indicate origin offset of (1,1).

for the binned scatterplot, thus making the loss explicit.

Figure 9 displays binned scatterplots with varying numbers of standard binned frequency groups. If we attempt to discern differences between similarly shaded tiles: it is trivial when only four shades exist, it becomes much more difficult at seven bins, and at ten frequency bins we are hardly able to discriminate better than in continuous shading. This aligns with Healey and Enn's theory on the number of discernible colors (Healey and Enns, 1999). Our exploration of frequency loss will focus on frequency binning with at most ten bins because above this we experience implicit frequency from perceptual bounds that are not well reflected in the explicitly defined frequency loss.

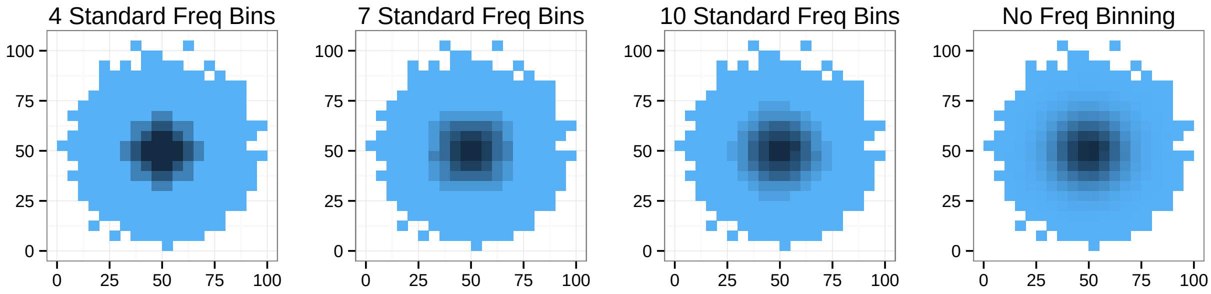


Figure 9: Binned scatterplots for the simulated bivarate normal data with varying numbers of standard binned frequency groups.

The loss of frequency information in frequency binning is dependent on the selection of a discrete color mapping, where the binning algorithm and number of frequency bins must be specified. Frequency loss is definitionally a decreasing function of the number of frequency bins for both standard and quantile frequency binning algorithms. The top row in Figure 10 displays the frequency loss from using both binning algorithms, with between one and ten frequency bins, from each set of simulated data. We first note the large difference in the magnitude of frequency losses based on the bivariate distribution; frequency losses are highest for the exponential data, lower for the normal data, and lowest for the uniform data. The decreasing frequency loss for both algorithms flattens out after the fourth frequency bin, each subsequent bin reducing the frequency loss less than the previous. Thus we should consider using between four and seven bin shades in order to reduce loss while also allowing for easy perception of frequency groups in the binned scatterplot.

Log transforming the frequencies prior to binning and using quantile based binning on the raw frequencies are two methods for dealing with the same problem for binned scatterplots: heavily right skewed bin counts where dense bins visually overshadow any structure in low density bins. It is strongly recommended to use the quantile-based algorithm to bin untransformed frequencies due to the improved handling of skewed bin count distributions while maintaining similar frequency loss to standard frequency binning when four to seven frequency bins are used, as seen in Figure 10. Frequency groups for quantile binning are invariant to the log transformation because a monotone transformation does not affect groupings based on quantiles, thus it is preferable leave the frequencies untransformed before quantile binning for better contextual interpretability.

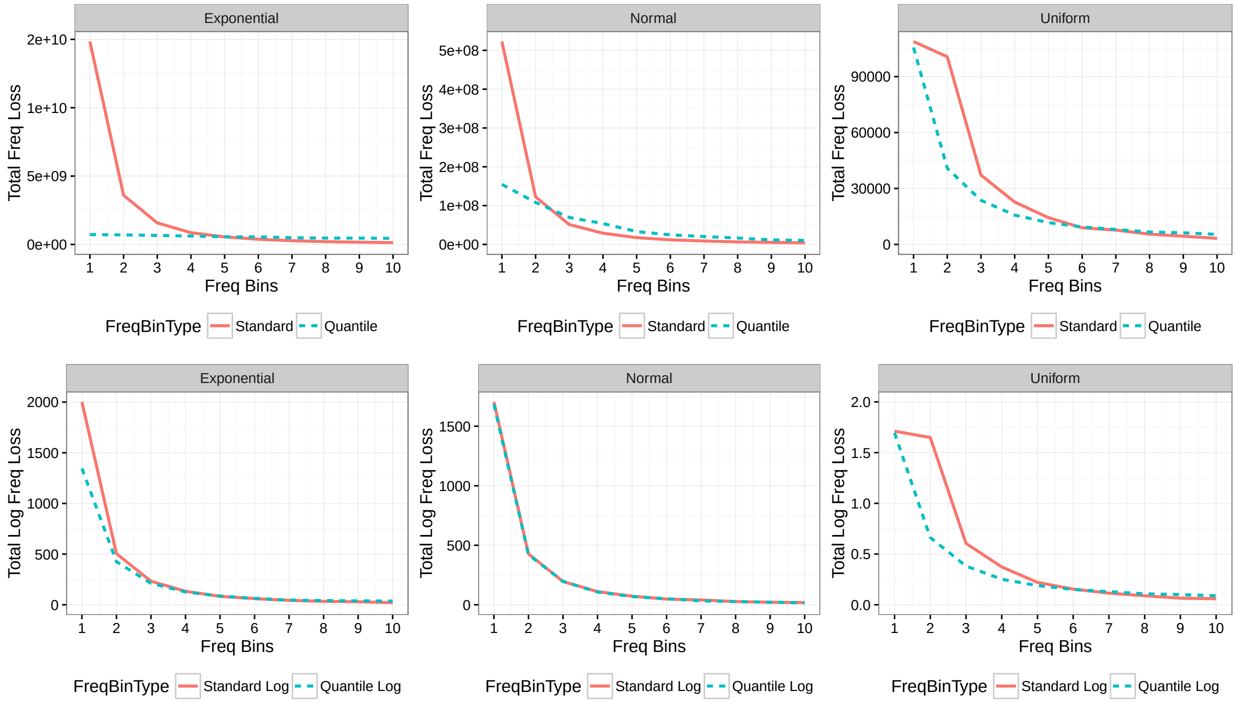


Figure 10: Lineplots for total frequency loss (top row) and total log frequency loss (bottom row) from standard and quantile binning of bin counts and log bin counts, respectively.

Due to the difference in scales between counts and log counts, the frequency loss and log frequency loss can not be directly compared. The bottom row of Figure 10 displays the log frequency loss from using standard and quantile algorithms for binning the *log* counts for bins from the same sets of simulated data. Log frequency binning behaved very similarly to standard frequency binning, where log frequency loss decreased as the number of frequency bins increased. The same advice to use between four and seven frequency bins also holds for shading a binned scatterplot based on log counts. Since the loss scales are not comparable, the choice is guided by desired interpretation. For standard log frequency binning, the shade is to be interpreted as an ordinal indicator based on equally spaced groupings of log bin frequencies; whereas for quantile frequency binning the shade denotes groups based on frequency quantiles. This is analogous to the difference in interpreting a histogram of log transformed data and a boxplot of untransformed data in univariate visualization. Both shading schemes are effective at reducing the visual impact of the highest density bins near the center of each plot, allowing for the differences in the surrounding bin frequencies to be emphasized.

5.3 Binning Loss in Baseball Data

We now revisit the baseball data used earlier and construct binned scatterplots using the loss-based specification principles developed in the previous sections. In the spatial binning step we learned three rules of thumb: (1) use smaller bins to achieve lower loss of spatial information, (2) standard binning has less loss than random binning but we need to be careful to choose bin dimensions that align with data resolution, and (3) offset the origin by a factor of half the data resolution to better center points within bins.

For the baseball data this leads us to specify small standard rectangular bins for data that is recorded with a one game by one strike-out data resolution. The data has only 42,583 observations, so we can use minimal binning to match the data resolution without taking an inordinate amount of computation time. While minimal binning eliminates spatial loss, the minuscule physical size of the resulting tiles makes it difficult to see the frequency shade; Therefore slightly larger two game by ten strikeout bins are used, providing approximately 50 bins per dimension. The origin offset of $(0.5, -0.5)$ is used to shift the binning back by a half game and down by one half strikeout to adjust for the data resolution. This offset for reduces the spatial loss by approximately 8% for this bin size.

The leftmost plot in Figure 11 displays the binned scatterplot with a continuous shading of raw bin frequencies. The most striking feature in the frequency distribution is the dark spot in the bottom right of the plot representing a large number of pitchers that played very few games and had very few strikeouts. It is nearly impossible to distinguish the density structure across the remaining bins, representing better pitchers who played many games and earned many strikeouts. In this case there are the two frequency binning approaches from Section 5.2 that we can employ to deal with this skewness of the distribution of bin

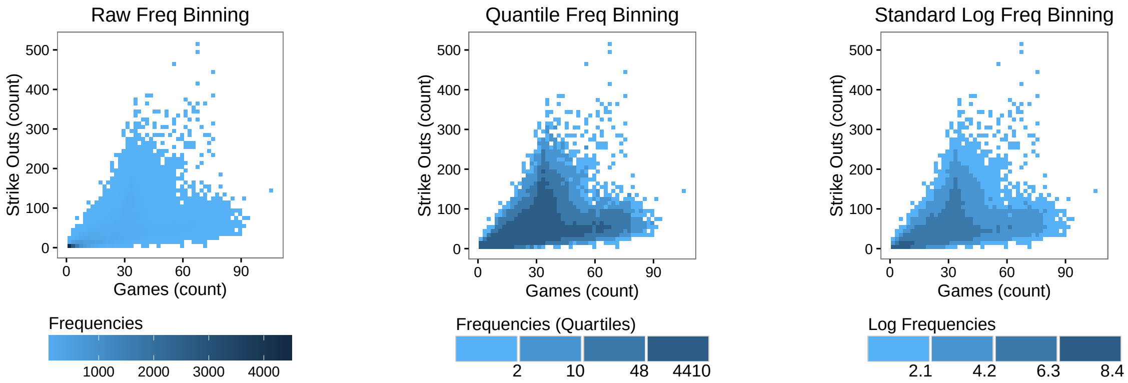


Figure 11: Binned scatterplots for games versus strikeouts.

counts. Quantile frequency binning using four shade groups (center plot of Figure 11) allows us to visualizes the quartiles of the bin densities. Alternatively, standard log frequency binning using four shade groups (rightmost plot of Figure 11) uses the log transformation to diminish the visual impact of high density bins. Each frequency binning approach adds a layer of complexity to interpreting the shade, however they both effectively emphasize the forked ridges in the density structure.

6 Conclusions and Future Work

Large bivariate data sets are very difficult to visualize in raw form, with the implicit loss of information due to over-plotting of points. Binning allows for the visualization and manipulation of large data sets, and easily translates into binned scatterplots which are more appropriate for the human visual system. Reducing the data for binned scatterplots has distinct computational and visual advantages, however the aggregation comes at the cost of losing precision in spatial and frequency information. We have introduced spatial and frequency loss functions to explicitly quantify the lost information in the construction of the binned scatterplot. This offers an optimizing criterion to guide binning specification and a foundation for future work on properties of a binned scatterplot as a visual estimator of the bivariate density – such as unbiasedness, consistency and efficiency.

We have presented two algorithms for spatially binning data points; standard and random rectangular binning algorithms. Our proposed random binning algorithm displayed strong advantage of avoiding the problem of artificial stripes that occur when data recorded to a coarse resolution are binned using a bin width that was a non-integer multiple of the data resolution. However, the standard binning algorithm is superior due to lower spatial loss. For data with a coarse resolution (α_x units in the X dimension and α_y units in the Y dimension) artificial stripes in the standard binning process can be avoided, if bin dimensions are chosen as integer multiples of α_x and α_y. We were also able to show through simulation that a reasonable default for the binning uses an origin offset of $(\alpha_x/2, \alpha_y/2)$ because it resulted in minimal or near minimal spatial loss for symmetric data and performed well for heavily skewed data.

Spatial binning with smaller bin dimensions will lead to lower spatial losses; however, finer binning requires more processing time and does not highlight large scale density structure. It is left to the plot designer to decide how much spatial information they are willing to sacrifice in order to simplify the display of density structure.

Due to imperfect human perception, we may elect to use frequency binning to discretize the shading scale with the goal to make the loss of frequency information more explicit. If we aim to have highly accurate perceptual mapping, it is recommended to use at most between four and seven distinct shades. This is done to minimize the frequency loss within the bounds of human perceptual ability to distinguish multiple shades simultaneously. Using quantile frequency binning or standard log frequency binning are shown to be reasonable methods – with slight differences of interpretability – for handling situations with heavily skewed bin counts. Frequency binning to discretize the shading scale allows us to explicitly quantify the loss of frequency information, however future perception research is needed to quantify the implicit frequency loss in reading binned scatterplots using continuous shading and the impact of contrasting effects.

Future implementations of software should consider these loss-related findings in graphical tools for constructing binned scatterplots. This research provides suggestions for reasonable default settings of binning parameters that maintain spatial and frequency information and lead to desirable visual properties in the binned scatterplot.

A Appendix for Origin Offset Proof

The following is proof that when univariate data is uniformly distributed at resolution α and standard rectangular binning is used with binwidths that are a scalar multiple of the data resolution (i.e. $\omega = k\alpha$ for some $k \in \{1, 2, \dots\}$), then spatial loss is minimized by setting the binning origin to $\alpha/2$ units below the minimum data value; set $\beta = x_{(1)} - \alpha/2$.

Let $x_1, x_2, \dots, x_k \in \mathbb{R}$ represent the values in a single bin such that $x_{i+1} = x_i + \alpha$ for some constant $\alpha \in \mathbb{R}$. Thus $x_j = x_1 + (j-1)\alpha$.

Suppose then that we bin the data using standard rectangular binning with origin, $\beta = x_1 - \theta$, and binwidth ω; where θ is the *origin offset* from the data. Thus $b(x_j) = \beta + \omega/2 = (x_1 - \theta) + (k\alpha/2)$

Spatial Loss, $L^S = \sum_{i=1}^{k} ||x_i - b_(x_i)||$ is definitionally minimized when $b_(x_i)$ is the *geometric median*. The geometric median for $x_1, \dots, x_k = Q_x(.5) = (x_{\lceil \frac{k+1}{2} \rceil} + x_{\lfloor \frac{k+1}{2} \rfloor})/2$, where $Q_x(\cdot)$ is the empirical quantile function.

Thus the optimal offset is the θ such that

$b(x_i) = Q_x(.5)$

$\Rightarrow (x_1 - \theta) + (k\alpha/2) = (x_{\lceil \frac{k+1}{2} \rceil} + x_{\lfloor \frac{k+1}{2} \rfloor})/2$

$\Rightarrow 2x_1 - 2\theta + k\alpha = (x_1 + (\lceil \frac{k+1}{2} \rceil - 1)\alpha) + (x_1 + (\lfloor \frac{k+1}{2} \rfloor - 1)\alpha)$

$\Rightarrow -2\theta + k\alpha = (\lceil \frac{k+1}{2} \rceil - 1)\alpha + (\lfloor \frac{k+1}{2} \rfloor - 1)\alpha$

$\Rightarrow -2\theta + k\alpha = ((k+1) - 2)\alpha$

$\Rightarrow -2\theta = -\alpha$

$\Rightarrow \theta = \alpha/2$

Thus the optimal offset for reducing spatial loss in this scenario is $\theta = \alpha/2$. This result holds for data that is symmetrically distributed within the bin since the median will not change. It extends to multiple contiguous bins with resolution α data that has symmetrically distributed data withing each bin.

If the same conditions are extended to the two dimensional case, then the origin for minimal spatial loss is at $(x_{(1)} - \alpha_x/2, y_{(1)} - \alpha_y/2)$ where α_x and α_y are the data resolution for each dimension, respectively.

References

Adelson, E. H. (1993), "Perceptual organization and the judgment of brightness," *Science*, 262, 2042–2044.

Bartleson, C. and Breneman, E. (1967), "Brightness perception in complex fields," *Josa*, 57, 953–957.

Carr, D., Littlefield, R., Nicholson, W., and Littlefield, J. (1987), "Scatterplot Matrix Techniques for Large N," *Journal of the American Statistical Association*, 82, 424–436.

Fairchild, M. D. (2013), *Color appearance models*, John Wiley & Sons.

Friedman, J. H. (1997), "Data mining and statistics: What's the connection," in *Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics.*

Goldstein, E. B. (2007), *Sensation & Perception*, Belmont, CA: Thomason Wadsworth.

Hao, M. C., Dayal, U., Sharma, R. K., Keim, D. A., and Janetzko, H. (2010), "Visual Analytics of Large Multidimensional Data Using Variable Binned Scatter Plots," in *IS&T/SPIE Electronic Imaging*, International Society for Optics and Photonics, pp. 753006–753006.

Healey, C. G. and Enns, J. T. (1999), "Large Datasets at a glance: combining textures and colors in scientific visualization," *IEEE Transactions on Visualization and Computer Graphics*, 5, 145–167.

Hermann, L. (1870), "Eine Erscheinung simultanen Contrastes," *Archiv für die gesamte Physiologie des Menschen und der Tiere*, 3, 13–15.

Hyndman, R. J. and Fan, Y. (1996), "Sample quantiles in statistical packages," *The American Statistician*, 50, 361–365.

Jacobs, A. (2009), "The Pathologies of Big Data," *Communications of the ACM*, 52, 36–44.

Janetzko, H., Hao, M., Mittelstadt, S., Dayal, U., and Keim, D. (2013), "Enhancing Scatter Plots Using Ellipsoid Pixel Placement and Shading," in *2013 46th Hawaii International Conference on System Sciences (HICSS)*, pp. 1–10.

Keim, D., Hao, M., Dayal, U., Janetzko, H., and Bak, P. (2010), "Generalized Scatter Plots," *Information Visualization*, 9, 301–311.

Leong, J. (2006), `http://hypertextbook.com/facts/2006/JenniferLeong.shtml`, accessed: 12/17/2015.

Liu, Z., Jiang, B., and Heer, J. (2013), "*imMens*: Real-Time Visual Querying of Big Data," in *Eurographics Conference on Visualization (EuroVis)*, International Society for Optics and Photonics, vol. 32.

Pearson, K. (1895), "Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material," *Philosophical Transactions of the Royal Society of London*, 186, 343–414.

Playfair, W. (1786), *Commercial and Political Atlas*, London.

Playfair, W., Wainer, H., and Spence, I. (2005), *Playfair's Commercial and Political Atlas and Statistical Breviary*, Cambridge University Press.

Scott, D. (1979), "On Optimal and Data-Based Histograms," *Biometrika*, 66, 605–610.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice and Visualization*, John Wiley & Sons.

Spillmann, L. (1994), "The Hermann Grid Illusion: a Tool for Studying Human Perceptive Field Organization," *Perception*, 23, 691–708.

Sturges, H. A. (1926), "The choice of a class interval," *Journal of the American Statistical Association*, 21, 65–66.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Lebanon, IN: Addison Wesley.

Unwin, A., Theus, M., and Hofmann, H. (2006a), *Graphics of Large Datasets*, New York: Springer.

— (2006b), *Graphics of Large Datasets*, New York: Springer.

Wickham, H. (2013), "Bin-Summarize-Smooth: A Framework for Visualising Large Data," Tech. rep.

Wilkinson, L., Anand, A., and Grossman, R. L. (2005), "Graph-Theoretic Scagnostics." in *INFOVIS*, vol. 5, p. 21.

Wilkinson, L., Anand, A., and Urbanek, M. S. (2015), "Package 'scagnostics'," .

Wilkinson, L. and Wills, G. (2008), "Scagnostics distributions," *Journal of Computational and Graphical Statistics*, 17, 473–491.