the plot, the clutter, the sampling and its lens: occlusion measures for automatic clutter reduction

Geoffrey Ellis
Computing Department
Lancaster University
Lancaster, LA1 4YW, UK
+44 (0)1524 510340
g.ellis@comp.lancs.ac.uk

Alan Dix
Computing Department
Lancaster University
Lancaster, LA1 4YW, UK
+44 (0)1524 510319
alan@hcibook.com

Abstract

Previous work has demonstrated the use of random sampling in visualising large data sets and the practicality of a sampling lens in enabling focus+context viewing. Autosampling was proposed as a mechanism to maintain constant density within the lens without user intervention. However, this requires rapid calculation of density or clutter. This paper defines clutter in terms of the occlusion of plotted points and evaluates three possible occlusion metrics that can be used with parallel coordinate plots. An empirical study showed the relationship between these metrics was independent of location and could be explained with a surprisingly simple probabilistic model.

Keywords

Sampling, random sampling, lens, clutter, occlusion, density reduction, overplotting, information visualisation.

1. Introduction

We have proposed random sampling [4, 5] as an effective technique for reducing density in overcrowded displays. If there is too much data to fit on the screen, taking a random sample of the data, that will fit, not only removes overlapping data items and clutter but it also preserves any trends or patterns that exist in the data. Unlike other clutter reduction techniques, with sampling, the user does not have to decide which data to remove and the view remains spatially undistorted.

Recent work [6] demonstrated that the Sampling Lens allows sub-sampling in areas of high density whilst retaining a higher-sampling rate over a visualisation as a whole. However, as one moves from high density regions to less heavily plotted regions of the visualisation, the sub-sampling rate typically needs to be changed either manually or automatically. In order to implement autosampling we need (a) an effective measure of 'clutter' or 'density', and (b) an efficient way of calculating the measure.

In this paper we address the first issue by looking at several potential metrics to measure occlusion, which is essentially the amount of overlapping lines in parallel coordinate plots. We decided to apply autosampling on parallel coordinates as we have found them to be a demanding visualisation for sampling.

In Section 2, we give an overview of the Sampling Lens and autosampling and also consider some related literature. We describe our experimental platform and the dataset in Section 3. In Section 4 we compare the metrics through an empirical study and a theoretical model.

2. Sampling and the Sampling Lens

Sampling has been found to be very effective in clutter reduction for various types of visualisations that require individual data items or attributes to be represented on the display.

The user can adjust the sampling level to reduce the data density of a visualisation, consequently revealing features that are otherwise hidden in the dense regions. The density across visualisations is usually non-uniform; as a result, the low sampling rate required to investigate denser regions can make the data in less dense regions 'disappear'. A way round this problem is to adjust the sampling rate for different areas of the screen [1]. However, our contribution to resolving this issue is the Sampling Lens [6] – a moveable region with its own sampling control, to deal with this issue. This follows a tradition of visualisation 'lenses' [2] that apply transformations or add information to the area under focus.

The Sampling Lens sub-samples the points within its lens region (see Figure 1), which allows the user to investigate dense regions of a plot by reducing the lens sampling rate to an appropriate level, thus revealing interesting patterns and trends whilst still retaining the context of the lens region within the overall plot. A more detailed description of Sampling Lens can be found in [6].

2.1 Auto-sampling

The manual adjustment of the lens sampling rate was found to be particularly tiring for the user. Instead the system could itself set the lens sampling rate, based on a measurement of the density of the points or lines within the lens. This was not an issue with scatterplots as only the number of data items at each display point had to be counted to measure the density.

The density estimation is however more challenging with parallel coordinate plots as the density of overlapping lines need to be measured. We have found very little work which defines a measure of density or clutter for overlapping lines.

2.2 Related work

Well known techniques for clutter reduction include filtering, distortion, clustering/aggregation, reordering, space-filling, constant density and sampling.

Rosenholtz et. al [9] provide a useful discussion of display clutter but notes that most of the metrics have problems and few have

Conference'04, Month 1–2, 2004, City, State, Country.

been implemented so far. They also describe a new measure of clutter, based on predicting the level of feature congestion in maps, using image values such as luminance and colour contrast; however this is not readily applicable to either parallel coordinate or scatter plots.

Tufte's [10] 'data ink ratio' and Frank and Timpf [7] 'ink per unit area' give a measure of crowdedness for traditional graphs and maps but they do not include any notion of hidden-ness.

Brath [3] in his "Metrics for Effective Information Visualization" gives several metrics, the relevant ones for 2D plots being:

data density = no. of data points ⁄ no. of available pixels
occlusion percentage = no. of points completely obscured ⁄ no. of points.

Bertini and Santucci's work on reducing clutter in scatterplots give some 'quality measures' that include a calculation of the number of collisions or overlapping points [1]. They divide the plot into small squares and use sampling to reduce the collisions in each square, whilst attempting to preserve the relative data density between all the squares in the plot. This is similar to the non-uniform sampling suggestion we proposed in [5].

Miller and Wegman [8] consider the plot densities for lines constructed in parallel coordinates. They study theoretical density plots rather than individual data points and use these to produce visualisations of density functions over high dimensional spaces. While the early parts of their work have some mathematical errors, they nevertheless manage to use their density plots to explore the properties of very hard to visualise multidimensional functions. Note this work does not measure clutter per se, indeed with a theoretical probability distribution they effectively have an infinite number of infinitely thin lines, but plot density is clearly closely related to clutter.

3. Dataset and experimental platform

The Sampling Lens application has been implemented in Java using the InfoVis Toolkit [http://ivtk.sourceforge.net/]. A significant amount of code has been added to provide the sampling lens functionality in both parallel coordinate plots and scatterplots. The experiments use an instrumented version of the Sampling Lens, which collects statistics about the measures being investigated.

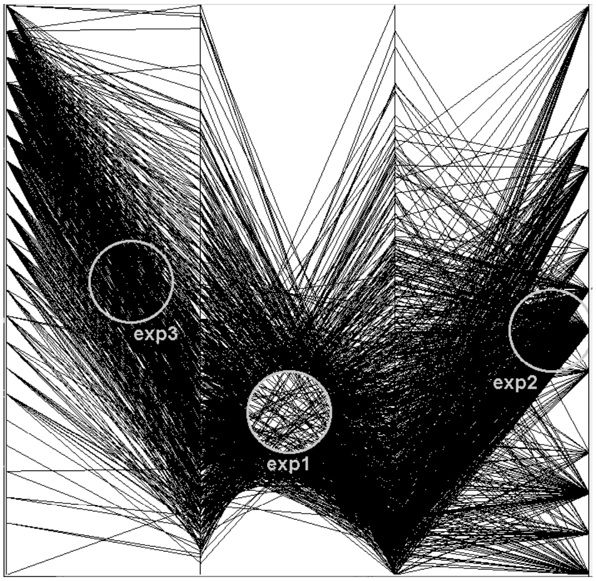

Figure 1. Parallel coordinate plot using 1K car dataset (labels and lens positions for exp2 & exp3 are superimposed)

The data used in most of the experiments is from the Portland cars dataset [31/3/05 http://www.cars.com]. The 5850 records contain details of cars for sale within 40 miles of Portland, Oregon. Figure 1 shows a screen shot of the parallel coordinate visualisation based on 1000 records of the cars dataset. Note that we have made no attempt to re-organise the attributes to simplify the plot.

The lens positions for the main experiments (exp1, exp2, exp3) referenced in this paper are shown in Figure 1. These positions were chosen to exemplify different patterns of lines crossing the lens, as illustrated in Figure 2.

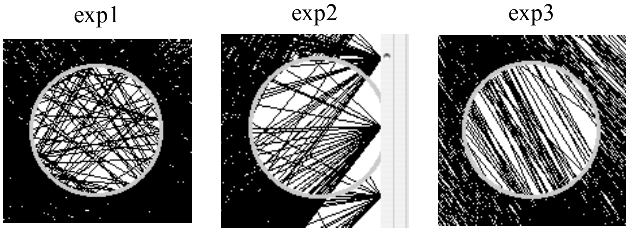

Figure 2. Lines within the lens at a 10% lens sampling rate

4. Metrics for clutter and density

From the literature review, it is clear that there is no commonly agreed measure for display density or clutter. Different things have an effect on perceived clutter, such as hidden or partially hidden screen objects, the closeness of adjacent objects and the merging of different coloured objects. In order to have a computationally tractable measure, we adopt fairly simple measures based on hidden points, as the important question for the user is "how much of the data cannot be seen?" Depending on the visualisation, a drawn object may be a single pixel point, point symbol, glyph, line or some text; all, apart from the first, may result in many pixels being plotted on the screen. In this investigation, we will only be considering the hidden or occluded objects and formulating this in terms of screen pixels.

4.1 Defining a measure of occlusion

For a given screen region (in particular the interior of the sampling lens), assuming S is the total number of available pixels and M is the number of plotted data points.

So, if the lens is circular with radius R pixels,

$$S = \pi R^2$$

and if there are L lines crossing the lens with an average pixels per line (ppl), then

$$M = ppl * L.$$

Note that, in general, M is not the number of actual pixels with points plotted on them as some points will be overplotted on the same pixel, so the number of plotted pixels is usually less than the number of plotted points.

We then define the following raw values from which we will obtain occlusion measures:

M_1 – number of plotted points on their own pixel
M_n – number of plotted points sharing a pixel
S_0 – number of empty pixels
S_1 – number of pixels with 1 plotted point (same as $M_{1)}$
S_n – number of pixels with more than 1 plotted point

Note that $M = M_1 + M_n$ and $S = S_0 + S_1 + S_n$ and always $M_1 = S_1$, but $M_n \geq 2\ S_n$ as each overplotted pixel contains two or more overplotted points.

Figure 3 shows an example of these values for a simple 3x3 plot.

An example of a 3x3 pixel section of the screen with a horizontal and vertical line crossing at the centre pixel.

$M = 6$, $S = 9$
$M_1 = 4$, $M_n = 2$
$S_0 = 4$, $S_1 = 4$ and $S_n = 1$

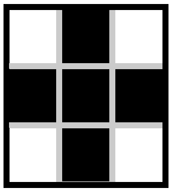

Figure 3. Example of overplotting

We can use these raw values to define three potential measures of occlusion: overplotted%, overcrowded% and hidden%. These are described with their formulae in Table 1. Note that in the example given in Figure 3, overplotted% = 20, overcrowded% ≈ 33 and hidden% ≈ 17.

Table 1. Definition of the occlusion measures

overplotted% $100 * S_n / (S_1 + S_n)$	The percentage of pixels with more than 1 plotted point. The range is 1 (all plotted points on their own) to 100 (no single plotted points)
overcrowded% $100 * M_n / (M)$	The percentage of plotted points that are in pixels with more than 1 plotted point. The range is 1 (all plotted points on their own) to 100 (no single plotted points)
hidden% $100 * (M_n - S_n) / (M)$	The percentage of plotted points that are hidden from view due to being overplotted. Note that pixels with more than 1 plotted point will be showing the top plotted point. The range is 0 to just less than 100, depending on the number of pixels.

These three occlusion measures are all close to Brath's measure:

number of points completely obscured / number of points

Although it is not clear whether 'points' means data points or actual plotted pixels and whether obscured means hidden; hence our desire to be precise in our definitions.

All three occlusion measures have some level of face validity and we find the first overplotted% most intuitive as it refers to the pixels visible to the user. However, they can vary substantially in theory. For example, if the display consisted of two identical lines (hence overplotted), then the overplotted% and overcrowded% values are both 100, whereas hidden% = 50. But if there are nine identical lines and one other non-intersecting line, the overplotted% value becomes 50 whereas overcrowded% = 90 and hidden% = 80.

In order to examine whether this is an issue in practice, we compared these measures empirically.

4.2 Empirical results

To investigate the relationship between the proposed occlusion measures, the raw values M_1, M_n etc. defined in section 4.1 need to be determined. This is achieved by rasterising the parallel coordinate lines to a pixel grid and counting how many points are plotted on each pixel. Figure 4 shows the calculated occlusion values over a range of sampling rates for two of the experiments. For each measure, there appears to be a definite trend with hidden% giving the lowest estimate of occlusion over the range of sampling rates and overcrowded% giving the highest estimate.

The computation of the occlusion measures is based on the number of pixels and the number of plotted points (see Table 1). As the lens size is the same in both exp1 and exp3, the number of pixels is constant, however the number of lines within the lens (and subsequent number of plotted points) is probably different due to the change in lens position between the experiments. To account for this, the graph was re-plotted using the number of plotted points as the x-axis. The lines for each experiment (e.g. overplotted% for exp1 and exp3) were now found to be coincident. This suggests that there is a fixed relationship between the equations for calculating the three measures. In order to verify this, results from three other lens positions were plotted on the graph and the lines for each measure were again found to be coincident.

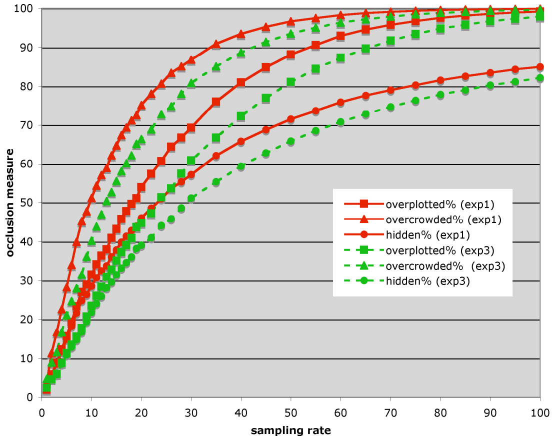

Figure 4. Occlusion measures vs. sampling rate

Figure 5 shows the same data as plotted in Figure 4, but this time the occlusion measures are normalised against the overplotted% values (i.e. the overplotted% value as the x-axis against occlusion measures on y-axis). The overplotted% line is straight and although there is a substantial difference between the measures, the relationship to overplotted% is the same independent of the chosen lens position. In particular, the relationship between all three occlusion measures is monotonic. This implies that:

(i) if we have an estimate or a calculated value for any one measure, we can derive the other two, or

(ii) if we have a 'target occlusion' slider, then it somehow does not matter which measure we use, as the legends on the slider would just differ slightly. Furthermore, if the user fixes the slider at any position, all three measures would be fixed as the sampling lens is moved over the screen.

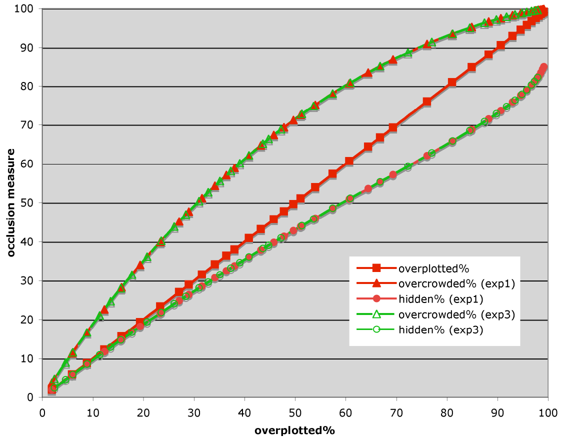

Figure 5. Normalised occlusion measures

4.3 Theoretical model

Figure 5 shows an entirely empirical relationship between the three occlusion measures. However, this empirical relationship corresponds closely to a very simple model. Imagine randomly plotting the M points over S' pixels (not necessarily all S pixels). Assuming that M and S' are quite large (so that combinatorics approximate to exponentials) and λ is defined as the ratio M/S', we can derive the expected values for the raw measures as:

$$\begin{aligned} E(M_1) &= M e^{-\lambda} \\ E(M_n) &= M(1 - e^{-\lambda}) \\ E(S_0) &= S e^{-\lambda} \\ E(S_1) &= M e^{-\lambda} = S \lambda e^{-\lambda} \\ E(S_n) &= S(1 - (1+\lambda) e^{-\lambda}) \end{aligned}$$

Using the above equations, we can calculate the expected overplotted%, overcrowded% and hidden% values as:

$$\begin{aligned} \text{overplotted\%} &= 100 (1 - (1+\lambda) e^{-\lambda}) / (1 - e^{-\lambda}) \\ \text{overcrowded\%} &= 100 (1 - e^{-\lambda}) \\ \text{hidden\%} &= 100 (1 - (1 - e^{-\lambda}) / \lambda) \end{aligned}$$

Figure 6 shows these expected values normalised against overplotted%. Notice the similarity with Figure 5 – although the real parallel coordinates points are not randomly placed (indeed they are in lines!) the relationship between the different measures of occlusion is precisely the same in practice as this very simple theoretical model.

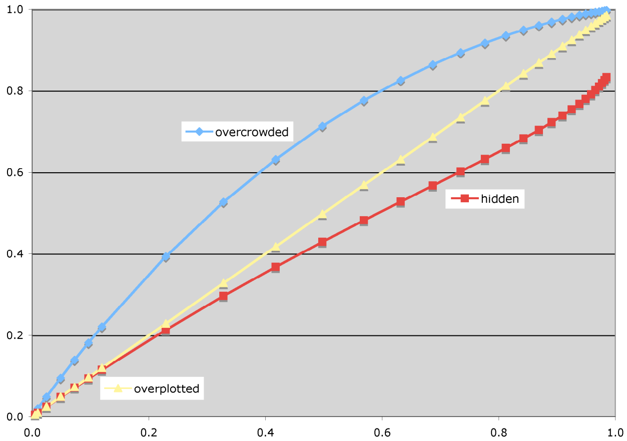

Figure 6. Theoretical curves for measures based on random point placement

Having looked at the different occlusion measures we now have to decide which one is the best. The overplotted% value is based on the number of occupied pixels (pixels positions), so in a sense it is a more viewer-centric measure. But overcrowded% and hidden% are based on the number of plotted pixels, hence they are more data-centric measures. Whilst all these measures are different, they are functionally related in practice, so a measure of any one is a measure of them all.

5. Conclusion

Parallel coordinate plots are a popular technique for exploring and hopefully gaining insight into, large multidimensional datasets. However, they often suffer from overcrowding due to large numbers of overlapping lines. Random sampling is an effective way of reducing this display clutter and when used within a moveable 'lens' region this provides a useful focus+ context tool, the Sampling Lens. Autosampling is necessary to prevent the user continually altering the sub-sampling rate of the lens, however this requires a method for estimating the density or clutter caused by overlapping lines.

We have defined three measures of occlusion and have demonstrated empirically that they are in fact related even though the chosen measures are based on different raw data values. Furthermore, we have developed a probalistic model and the theoretically calculated measures agree closely with the empirical results. This was quite unexpected, although we should point out that from the onset of our work with sampling, we have often used binomial and Poisson approximations to obtain order of magnitude ideas of behaviour.

Even if the lines were randomly scattered, the points on them would not be; they would be lines! Strangely this seems to be irrelevant to the bulk behaviour and it appears this simple (even simplistic) model is surprisingly good!

In related work, not reported here, we have already been looking at different algorithms to measure the occlusion of lines in parallel coordinate plots using the overplotted% metric, with the aim of producing efficient real-time adjustment of the sampling lens.

6. References

[1] Bertini, E. and Santucci, G. Improving 2D scatterplots effectiveness through sampling, displacement and user perception. *Proceedings of Information Visualisation 2005*, London, July 2005, IEEE

[2] Bier, E A., Stone, M C., Pier, K., Buxton, W., De Rose, T D. Toolglass and magic lenses: the see-through interface. *Proceedings of Computer Graphics and Interactive Techniques*, 1993, 73-80

[3] Brath, R. Concept Demonstration: Metrics for Effective Information Visualization. *Symposium on Information Visualization, Phoenix*, AZ, Oct 1997, IEEE, 108-111

[4] Dix, A. and Ellis, G.P. by chance: enhancing interaction with large data sets through statistical sampling. *Proceedings of the International Working Conference on Advanced Visual Interfaces*, L'Aquila, Italy, May 2002, ACM Press, 167-176

[5] Ellis, G.P. and Dix, A. Density control through random sampling : an architectural perspective. *Proceedings of*. *Information Visualisation 2002*, London, July 2002, IEEE, 82-90

[6] Ellis, G.P., Bertini, E. and Dix, A. The Sampling Lens:Making Sense of Saturated Visualisations . *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, Portland, USA, 2005, ACM Press, 1351-1354

[7] Frank, A.U. and Timpf, S. Multiple Representations for Cartographic Objects in a Multi-scale Tree – An Intelligent Graphical Zoom. *Computers and Graphics*, *18(6)*, 1994, 823-829

[8] Miller, J. and Wegman, E. Construction of line densities for parallel coordinate plots. *Computing and Graphics in Statistics*, IMA Volumes In Mathematics And Its Applications, 1992, Springer-Verlag, 107–123.

[9] Rosenholtz, R., Yuanzhen Li, Jonathan Mansfield, Zhenlan Jin. Feature Congestion: A Measure of Display Clutter. *Proceedings of the SIGCHI conference on Human factors in computing systems*, Apr 2005, ACM Press, 761-770

[10] Tufte, E.R. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 1983