Context-Rich Graphical Displays

David A. James
Bell Laboratories
Murray Hill, NJ 07974
dj@bell-labs.com

1 Introduction

In this short communication we briefly discuss the idea of context-rich graphical displays for exploratory data analysis. We illustrate this concept through two case studies and attempt to distill a handful of useful principles that may be used when analyzing data from other areas.

If we consider graphical techniques such as scatter plots, histograms, boxplots, quantile plots, etc., as context-free in the sense that they are applicable to data sets collected in any field of interest, context-rich then refers to graphical displays highly tailored to specific applications. This allows us, for instance, to augment familiar displays with information that is often implicit but not fully exploited by field investigators. By making explicit this information through graphics we are able to better articulate the analysis' goals, methods, conclusions, and limitations.

This idea has a long tradition. Many well-known displays are context-rich, for instance, Napoleon's failed Russian campaign [Tufte, 1983], Playfair's wheat price, wages, and the reigns of British kings from 1565 through 1821 [Tufte, 1983], cholera outbreak in London's 1850's [Cliff and Ord, 1981]; more recent examples include EVENTCHARTS [Goldman, 1992], Caveplots [Becker et al., 1994], SeeNet [Becker et al., 1991], etc. Extra information is effectively displayed in the context of the application in all of the above mentioned displays, e.g., weather hardships that Napoleon's army experienced, the steady economical improvement of the British empire, censoring in the case of EventCharts, network topology in the case of SeeNet, etc. See [Cleveland, 1993], [Tufte, 1983], [Tufte, 1990] for many more examples.

2 Case Studies

In the following case studies we show how graphical displays in combination with analysis of variance decompositions allows us to study the longitudinal and spatial dependence of covariates on manufacturing metrics. This approach is semi-parametric in that no longitudinal/spatial structure is imposed on the response or its covariates. In both examples, the response is measured at multiple points, in the first example along an optical fiber and on the surface of silicon wafers in the second example. The response is first smoothed and an analysis of variance model fitted at each location. We will fit linear models

$$\mathbf{Y} = \mathbf{X}\,\mathbf{B} + \mathbf{E}$$

where all elements in the model are matrices; $(y_{i,p})$ denotes the smoothed response for the $i'th$ observation at the $p'th$ position, $\mathbf{X}$ is a designed matrix corresponding to some parametrization of the covariates, $(\beta_{k,p})$ is the coefficient for the $k'th$ term at the $p'th$ position, and $(\epsilon_{i,p})$ is the residual from the $i'th$ observation at the $p'th$ position. From this process we get multiple sets of coefficients and effects (one set per location) that we display longitudinally along a fiber and spatially on a wafer.

2.1 Optical Fiber

Two important quality characteristics of the transmission along optical fiber cables are the power attenuation and dispersion as the light signals travel along the fiber. Their measurement is done with an optical time-domain reflectometer (OTDR) that sends laser pulses along the fiber and measures their back-scatter intensities at equally-spaced points. To measure the fiber attenuation at a single point on the fiber, multiple signals of specified width are sent and their backscatter averaged. This process is repeated for positions $p_1, p_2, \ldots, p_n$, to form a collection of power measurements (in dB) that traces the signal attenuation along the fiber. Figure 1 shows one such trace.

Although not distinguishable from Figure 1, the variance of the measured attenuation is largerly dependent on position (i.e., as the signal travels further into the fiber, the measurement error is known to increase). Various factors were known to have a possible effect on the variance, but the form of this dependence was not understood in the presence of manufacturing variability.

To characterize the OTDR measurement error along the fiber, a 24-run full factorial experiment was conducted varying the power of the input laser or "plugin" (low versus high), the width of individual light pulses (short, medium and long), and the number of pulses that are averaged at each point on the fiber (50, 100, 150, and 200). At each condition 100 replicates were taken and their loess [Cleveland, 1979] smoothed variance is shown in Figure 2.

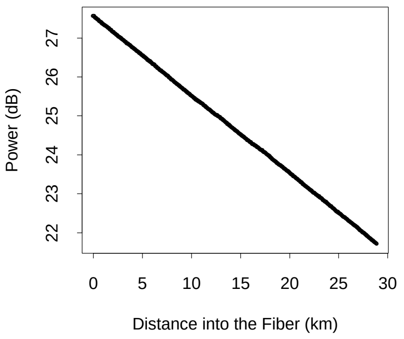

Figure 1: Power along an optical fiber.

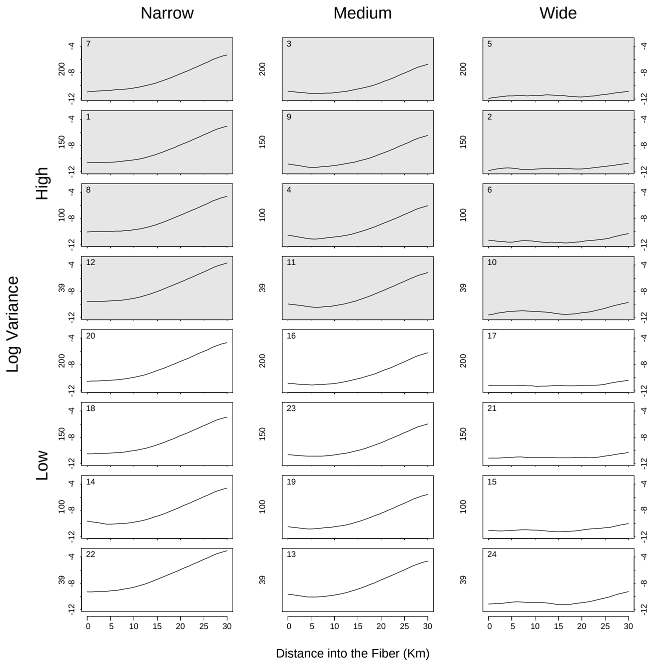

Figure 2: Designed settings from a full-factorial experiment. The top 12 shaded panels correspond to runs with high plugin and the bottom 12 panels with low plugin. Each column depicts the response at one pulse width. Each row shows the response for one level of number of averages: 200, 150, 100, and 39. The number in the upper left-hand corner of each panel indicates the order of experimentation.

Traditionally, experiments like this have been analyzed by first reducing the response (the variance traces) to scalars (e.g., means or medians variance). Then an analysis of variance would be conducted on these summaries. This approach ignores the possible dependence of the factors on positions along the fiber; if the effects are monotone and significantly large, adequate settings may be concluded, but the intrinsic variability along the fiber could not be assessed.

Instead, without too much extra work, we may estimate the effect of the covariates at fixed intervals along the fiber. Had we had information as to the functional form of the variance traces as the covariates vary, we would had modeled it directly. Since we did not, we fitted models at one-kilometer intervals to $\log(\hat{\sigma})$, and then collected the terms into traces to understand the effects' dependence on position. Figure 3 shows how the effects of the covariates vary along position on the fiber.

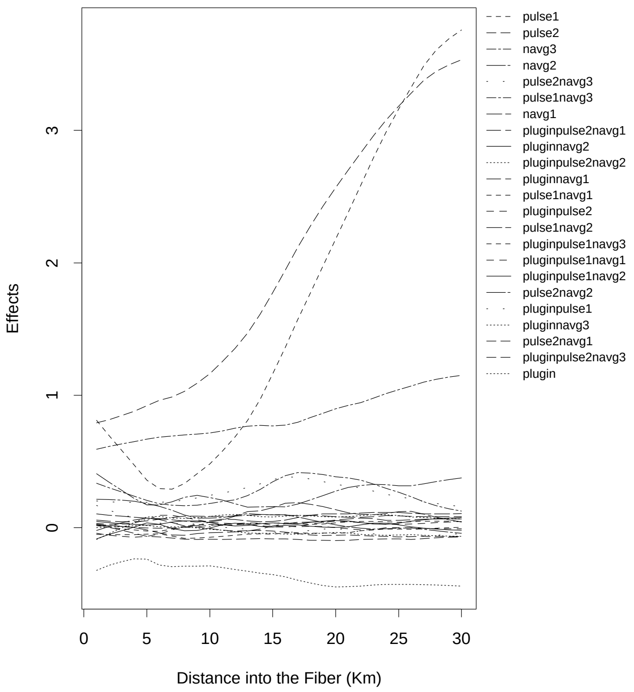

Figure 3: Effect traces from the full ANOVA model.

We see, for instance, that the two curves corresponding to the terms in pulse width contribute the most to measurement error, also apparent from Figure 1. Moreover, their contribution is nonlinearly increasing as a function of position. Similarly, we see that the "plugin" effect of input laser power (high better than low), is significant and somewhat uniform along the fiber. Finally, only one term from number of signal averages per measurement is shown to be significant and linearly increasing along the fiber: this term contrasts averaging 50 signals versus 100 or more. We concluded that 100 or more averages is adequate.

2.2 Integrated Circuits on a Wafer

Integrated-circuits or “chips” are manufactured on silicon wafers. These wafers are circular disks that may contain from 40 up to 800 or more chips. Before these wafers are cut and the microchips packaged, a battery of tests are applied to each chip to assess whether it is defective or not (for this presentation we will ignore the numerous outcomes that a chip may experience).

At each site on a wafer, we estimate its probability of being defective by using a smoother (we use a kernel smoother, but others may also be used) and then apply a sequence of transformations to stabilize their variance. Figure 4 shows one binary wafer where black squares denote defective chips and white squares denote non-defective chips; the wafer on the right shows a smoothed version of the binary wafer where dark areas denote regions of “high” defect probabilities (high with respect the overall proportion of defective chips.)

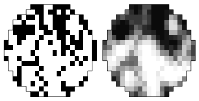

Figure 4: A binary wafer and its smoothed version.

We then may want to relate these probabilities to manufacturing parameters, for instance, through designed experiments.

Traditionally, engineers have summarized each wafer by its yield (i.e., the proportion of good chips) and performed data analysis on the yield alone. This approach, like the traditional approach seen in the fiber example, ignores the spatial structure of the response.

A designed experiment was conducted to study the effect of two factors on yield; the factors were a diffusion time and a coating step. Three diffusion times were varied, from short, medium, and long; two coatings were used, thin and thick.

In Figure 5 we show the interaction between the two factors by computing the average yield at each factor combination. We augment this well-known display with averaged wafers were the value at the *i'th* site represents the average defective proportion across wafers for the given factor combination.

Figure 5 suffices to determine the factor combination that maximizes yield (thick coating and short diffusion times). However, the engineers were also interested in untangling the effects of the factors on the wafer surface, thus an extra step was needed. By following the paradigm shown in the fiber example, we fitted an analysis of variance model at each site on the wafer. We collected the coefficients of coating and the linear term of diffusion from all ANOVA’s and displayed them as wafer objects in Figure 6.

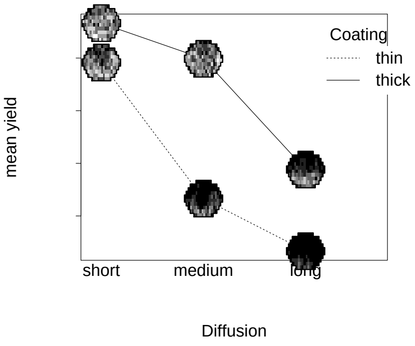

Figure 5: Interaction plot for the wafer experiment.

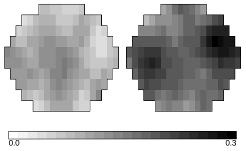

Figure 6: Coefficients for the coating term and for the linear term of diffusion time.

We then repeated this process for all elements of the models and display them as wafer objects in a familiar regression analysis summary table in Figure 7.

The first column in Figure 7 shows how the various coefficients vary on the wafer surface; the second column displays the standard error for each coefficient and site; similarly the third column shows T-values for each site and coefficient; the last column shows p-values — light sites (close to zero) denote chips where the corresponding coefficient is significantly different from zero. The two wafers at the bottom show the estimated standard deviation (labeled `Root MSE`) and the percentage of variability that the model explains (labeled `R-squared`): we note that the model (i.e., changes in diffusion and coating) can explain a large fraction of the variability in yield. The intercept wafer indicates that yield is smallest at the top of the wafers across all experimental conditions.

The coefficient for coating (better seen in Figure 6) shows smaller yields in the center of the wafer, while diffusion time has large coefficients outside the center.

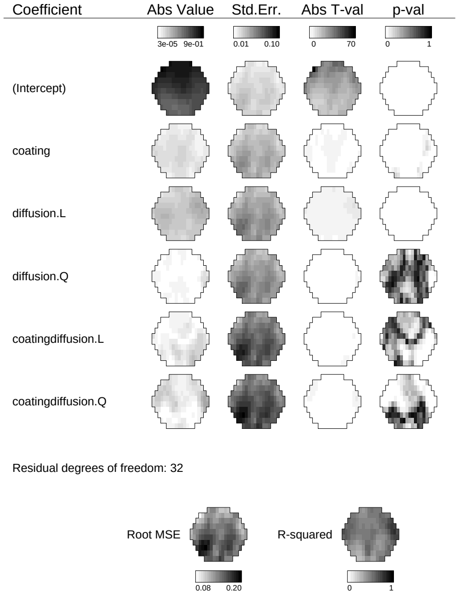

Figure 7: Coefficients from the ANOVA fit

Thus by carrying out the analysis of variance at each site we have been able to separate the effects of coating from those of diffusion time on the wafers.

Predictions and residuals may also be plotted as wafers for model diagnostics.

3 Discussion

As the examples illustrate, by incorporating intrinsic longitudinal/spatial structure into known graphical methods we begin to understand the variation in our data along those dimensions. Notice that in both examples we employed the analysis of variance decomposition as a means for exploratory data analysis rather than an inferential procedure, and then displayed its various components along the length of the fiber and over the surface of a wafer. It was through graphical displays that we assessed the extent of the longitudinal and spatial dependence of the responses. Alternative approaches that could be used include generalized additive models with varying coefficients, [Hastie and Tibshirani, 1990], [Hastie and Tibshirani, 1991], and loess models [Cleveland, 1979], [Cleveland et al., 1993], among others. In the case of wafers, formal spatial analysis techniques can be used to estimate the extent of spatial clustering and its relation to the covariates, for instance [Taam and Hamada, 1992], but the above graphical displays were more visually effective.

In both examples the models used did not include longitudinal/spatial components in the predictors because it was strongly felt that there was no prior knowledge that could reasonably describe the longitudinal/spatial structure in the response or the covariates. Moreover, by allowing the effects to freely vary over the fiber and over the wafer surface we guarded against most types of misspecification; yet the graphics we used to display the effects and coefficients effectively reveal their longitudinal and spatial dependence.

If the data are not collected through designed experiments, techniques such as principle components or hierarchical clustering may be appropriate. These are some of the multivariate techniques whose graphical displays can easily be augmented with similar symbols or glyphs, e.g., imagine a cluster dendrogram with wafers at the leaves.

Figures 5 is an example of this simple technique that uses glyphs to add contextual information to known graphical displays. Figure 8 below shows the distribution of planned yield for a set of wafers; three average wafers are superimposed to depict the spatial variation of yield for the lower 25% of the yield distribution the middle 50% and the top 25%.

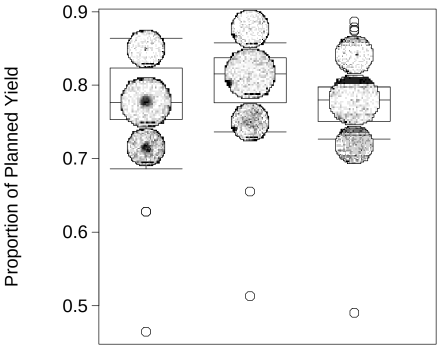

Figure 8: Boxplot of yield plus site-wise averages for the low 25%, middle 50% and top 25% of the yield distribution.

Looking at yield by itself clearly fails to convey systematic patterns of defectives that, if removed, would significantly improve production. The glyphs in this case are wafers, but we could imagine plotting boxplots, stars, thermometers, time-series or other glyphs on scatter plots or other displays as a means to show dependence (or lack thereof) among the variables.

4 Acknowledgments

The fiber study was joint work with Daryl Pregibon, who introduced me to the idea of "pasting" ANOVA models; the wafer case study was joint work with Daryl and Mark H. Hansen – I want to thank both very much.

References

[Becker et al., 1994] Becker, R. A., Clark, L. A., and Lambert, D. (1994). Cave plots: A graphical technique for comparing time series. *Journal of Computational and Graphical Statistics*, 3(3):277–284.

[Becker et al., 1991] Becker, R. A., Eick, S. G., and Wilks, A. R. (1991). Basics of network visualization. *IEEE Computer Graphics and Applications*, 11(3):12–14.

[Cleveland, 1979] Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scaterplots. *Journal of the American Statistical Association*, 74:829–836.

[Cleveland, 1993] Cleveland, W. S. (1993). *Visualizing Data.* Hobart Press, Summit, NJ.

[Cleveland et al., 1993] Cleveland, W. S., Mallows, C. L., and McRae, J. E. (1993). ATS methods: Nonparametric regression for non-gaussian data. *Journal of the American Statistical Association*, 88(423):821–835.

[Cliff and Ord, 1981] Cliff, A. D. and Ord, J. K. (1981). *Spatial Processes: Models and Applications.* Pion Limited, London, UK.

[Goldman, 1992] Goldman, A. I. (1992). Eventcharts: Visualizing survival and other timed-events data. *The American Statistician*, 46(1):13–18.

[Hastie and Tibshirani, 1991] Hastie, T. and Tibshirani, R. (1991). Varying-coefficient models. Technical memorandum, AT&T Bell Laboratories, Murray Hill, NJ.

[Hastie and Tibshirani, 1990] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

[Taam and Hamada, 1992] Taam, W. and Hamada, M. (1992). Detecting spatial effects from factorial experiments: An application from integrated-circuit manufacturing. *Technometrics*, 35:149–160.

[Tufte, 1983] Tufte, E. R. (1983). *The Visual Display of Quantitative Information.* Graphics Press, Cheshire, Connecticut.

[Tufte, 1990] Tufte, E. R. (1990). *Envisioning Information.* Graphics Press, Cheshire, Connecticut.