

Interactive and Dynamic Graphics in Statistical Consulting

Harold V. HENDERSON

Twelve years ago, Unwin wrote an article titled "How Interactive Graphics Will Revolutionize Statistical Practice." Well, has it? I think it has. At that time Data Desk was the prime commercially available package for interactive statistical graphics. Since then interactive and dynamic graphics have become increasingly available in many statistical and visualization packages, and some are even available, or easy to set up, in general spreadsheets like Excel. Examples of using interactive and dynamic graphics in ways that are useful in statistical consulting will be given. We'll gain insight into some old data and into some new data, including microarray data.

Key Words: Brushing; Data analysis; Data visualization; Exploratory data analysis; Interactive graphics; Linked scatterplots; Microarray data; Slicing; Statistical graphics.

1. INTRODUCTION

Twelve years ago, Unwin (1992) wrote an article titled "How Interactive Graphics Will Revolutionize Statistical Practice." He noted that "Interactive graphics have changed how statisticians explore data. They will also change how statisticians work and communicate with others." He used two well-known datasets to illustrate what interactive graphics can add to the interaction between statisticians and data-owners. He concluded that "Interactive graphics encourages statisticians to do more exploratory data analysis and to confirm results graphically. Interactive graphics involves the data owner more, brings him/her closer to the data and stimulates him/her to give more information. Interactive graphics provides the tools which support and encourage interaction between data-owners and statisticians. That will revolutionize statistical practice." Becker, Cleveland, and Wilks (1987) predicted that "in the future, dynamic graphical methods will be ubiquitous." In these two articles, the terms "interactive graphics" and "dynamic graphics" were essentially referring to the same

Harold Henderson is Statistician, AgResearch, Ruakura Research Centre, Private Bag 3012, Hamilton, New Zealand (E-mail: Harold.Henderson@AgResearch.co.nz).

©2004 American Statistical Association and the International Biometric Society
Journal of Agricultural, Biological, and Environmental Statistics, Volume 9, Number 4, Pages 402–431
DOI: 10.1198/108571104X15494

features, characterized by direct manipulation and instantaneous change.

Today, a search on Google (September 20, 2004) gave 97,500 hits for “dynamic graphics”, 76,000 for “interactive graphics”, and 610 with them both. Unwin (1999) proposed a basic set of requirements for interactive graphics software for exploratory data analysis: “To be fully effective it needs fast, flexible, and usable implementations of the following interactive tools: direct querying, zooming, rescaling, selection with linking, and the use of multiple views. In addition, it needs to support all these operations for selected subgroups. The interface must be consistent throughout and each tool or method must be fully integrated into the system.” He concludes that all of these tools “are valuable in discerning the information in data sets. They enable exploratory data analysis to be carried out in a flexible and insightful manner.” A useful distinction is that “interactive graphics” includes these basic interactive tools and “dynamic graphics” includes some additional tools like rotating plots and dynamically transformed axes. So I will refer to interactive and dynamic graphics.

Well, has interactive graphics revolutionized statistical practice? I think it has. At that time Data Desk was the prime commercially available package for interactive and dynamic graphics. Since then interactive and dynamic graphics have become increasingly available in many statistical and visualization packages, and some are even available, or easy to set up, in general spreadsheets like Excel. So today, in addition to Data Desk we have, for example, GGobi, SAS Insight, the stand-alone JMP from SAS, to mention just a few. The GGobi software is a data visualization system with state-of-the-art interactive and dynamic methods for the manipulation of views of data. A recent development is an R interface for GGobi so that GGobi can be run from R (and soon S-Plus). Cook, Swayne, Buja, and Hofmann are preparing a book on *Interactive and Dynamic Graphics for Data Analysis: With Examples Using R and GGobi*.

An early use of dynamic graphics was plotting power transformed data (Fowlkes 1971) where the power was changed by a knob. It is widely available now in statistical packages and is even easy to do in Excel, as is shown in Section 4, with scroll bars dynamically controlling the power for each axis.

Another early use of dynamic graphics was in rotating data. The first program for rotating data was PRIM-9 developed at AT&T Bell Labs by Fishkeller, Friedman, and Tukey (1974). PRIM was an acronym for Projection, Rotation, Isolation, and Masking, tools that were the basis for plot rotations. It was implemented on several million dollars' worth of computer. MacSpin (Donoho, Donoho, and Gasko 1985) brought 3D rotating plots to an inexpensive platform, the Macintosh. An early success of 3D rotating plots was for diabetes data explored using PRIM-9 and reported by Reaven and Miller (1979). But interactive graphics is much more than just 3D rotating plots or dynamic transformations, which are examples of dynamic graphics.

In making the case for statistical graphics, Scott (2003, p. 20) wrote that “Graphics are at the heart of modern statistical practice and discovery, not just to provide a summary of findings.” He continued (p. 21) “The area of most research and impact for exploratory graphics is in the field of dynamic graphics. . . . The tools of brushing and linking are critical to modern exploratory data analysis.” Plot brushing was initially developed as a way to work

with scatterplot matrices, but is now more generally available in any appropriate display. One of the most useful features of interactive graphics is this linking between different displays, whereby, selecting points in one display highlights the parts of other plots that display those points. You can then see highlighted points against the background of other points or when you turn on “hot selection” (as it is called in Data Desk) you see just the highlighted points. For example, you can select each bar of a bar chart in turn and observe the changes in graphics and analysis windows. Tukey (1973) called this successive viewing of the same display for groups “alternagraphic” viewing. The name has not caught on but the technique is extremely useful.

A complementary tool is trellis display (Becker, Cleveland, and Shyu 1996). Trellis displays present panels of the same display for different groups laid out in rows, columns, and pages. Each panel shows the relationship of certain variables conditioned on the values of other variables. Examples of these are among the displays that Tufte (1990) described as small multiples. Tufte (1990, p. 68) wrote, “Small multiples reveal, all at once, a scope of alternatives, a range of options.”

The power of interactive graphics is that it helps you discover unexpected patterns, structure, and relationships among several variables, by touching graphs and watching them change in response to your actions.

In recent years some interactive and dynamic graphics have even become available, or easy to set up, in Excel. The widespread use of Excel means that even without specialized software some interactive graphics are available to general users. In AgResearch, our scientists generally maintain their data in Excel and we encourage “good spreadsheet practice” and simple exploratory data analysis in Excel. For example, filters give interactive graphics by showing plots of just the data selected by the filter. Pivot tables and charts provide interactive tables and plots summarized by variables that can be dragged and dropped onto or off the table or graph. See the one-page “Spreadsheet Guidelines.doc” by my colleague, Neil Cox, at <http://www.agresearch.co.nz/downloads/graphics/>. It is surprising that Excel does not provide histograms or boxplots as standard features, although they can be set up with some effort. Even 3D rotating plots are easy to set up in Excel. For example, rotating around the y axis with transformed $x' = x \cos(\text{angle } y) + z \sin(\text{angle } y)$ and angle y linked to a scroll bar from the control toolbox. A 3D rotating plot of the diabetes data (“rotation diabetes.xls”) is at <http://www.agresearch.co.nz/downloads/graphics/>. Simulated manual plotting can be set up with scroll bars linked to lengths of dynamic ranges as in “abrasion.xls” at the same site.

2. USE IN STATISTICAL CONSULTING

*“I hear and I forget, I see and I remember,
I do and I understand.” —Confucius*

In the following sections, I use three datasets to illustrate examples of using interactive and dynamic graphics in ways that are useful in statistical consulting.

In statistical consulting it is often important (and rewarding) to gain insight into the structure of data. Sometimes the structure may not be explicit and patterns in the data or description of how the data were collected may suggest there is some structure to the data and suggest some questions to ask the data owner about the structure. This is now illustrated with a small old dataset where interactive graphics quickly reveals structure of the data that has been largely overlooked for nearly 60 years.

3. INTERACTIVE GRAPHICS REVEALS STRUCTURE IN ABRASION DATA

Nested structure in the well-known abrasion loss data (Davies 1947) is easily revealed by interactive graphics and appears to have been largely overlooked for nearly 60 years. Hand et al. (1994) list the data in *Handbook of Small Data Sets* and it has been widely used as an example of regression with two explanatory variables, for example: Chambers, Cleveland, Kleiner, and Tukey (1983); Cleveland (1985, pp. 213–218); and Cleveland (1993, chap. 4).

That the data has structure was alluded to when Davies (1947, pp. 118–119) introduced the data in a popular text to illustrate regression, noting that “Varying amounts and types of ingredients were used in the preparation of these samples in order to determine their effects on the properties of rubber.” In the third edition, Davies et al. (1957, p. 156) revised the comment to “Varying amounts and types of compounding material were used in the preparation of a number of specimens of this rubber and a wide range of physical properties was obtained.” But the data as listed in Davies (1947) and Davies et al. (1957) does not make the structure explicit. Only the three variables—loss, hardness, and tensile strength—are listed.

3.1 STRUCTURE DISCOVERED

The structure is easily revealed by interactive graphics. The use of interactive graphics is illustrated here using Data Desk but similar exploration could be done in other interactive graphics environments. It is always a challenge to show the process of interactive and dynamic graphics on paper, so you can download `abrasion.dsk` to interact with from <http://www.agresearch.co.nz/downloads/graphics/>. (You’ll also need Data Desk to use this file—you can download a trial version from http://www.datadesk.com/products/data_analysis/downloads/. I’ve used Data Desk version 6.1 (Velleman 1999).)

There appear to be five runs of increasing hardness on observation number in the first plot of Figure 1. The second plot is a graph of hardness on strength, the design space. Brushing points in the first plot highlights those points in the second plot. The remaining pairs of plots in Figure 1 show a sequence of dynamic plots brushing through runs of hardness with the selected points displayed in the plot of hardness on strength, labeled by observation number and the points in a run joined by a line in observation number order. The colors and lines are added dynamically.

There appear to be five “types” of compounding material and different colors and lines have been given dynamically for each “type.” Points are labeled by observation number. Observation 1 appears remote in the scatterplot of hardness on strength, but does not look out of place in the timeline of hardness, so we separate observation 1 from the group of observations 2 to 8. Although the structure is not explicit in the data listed in Davies (1947) and the third edition Davies et al. (1957), the data order appears to be in *amount* within *type* order.

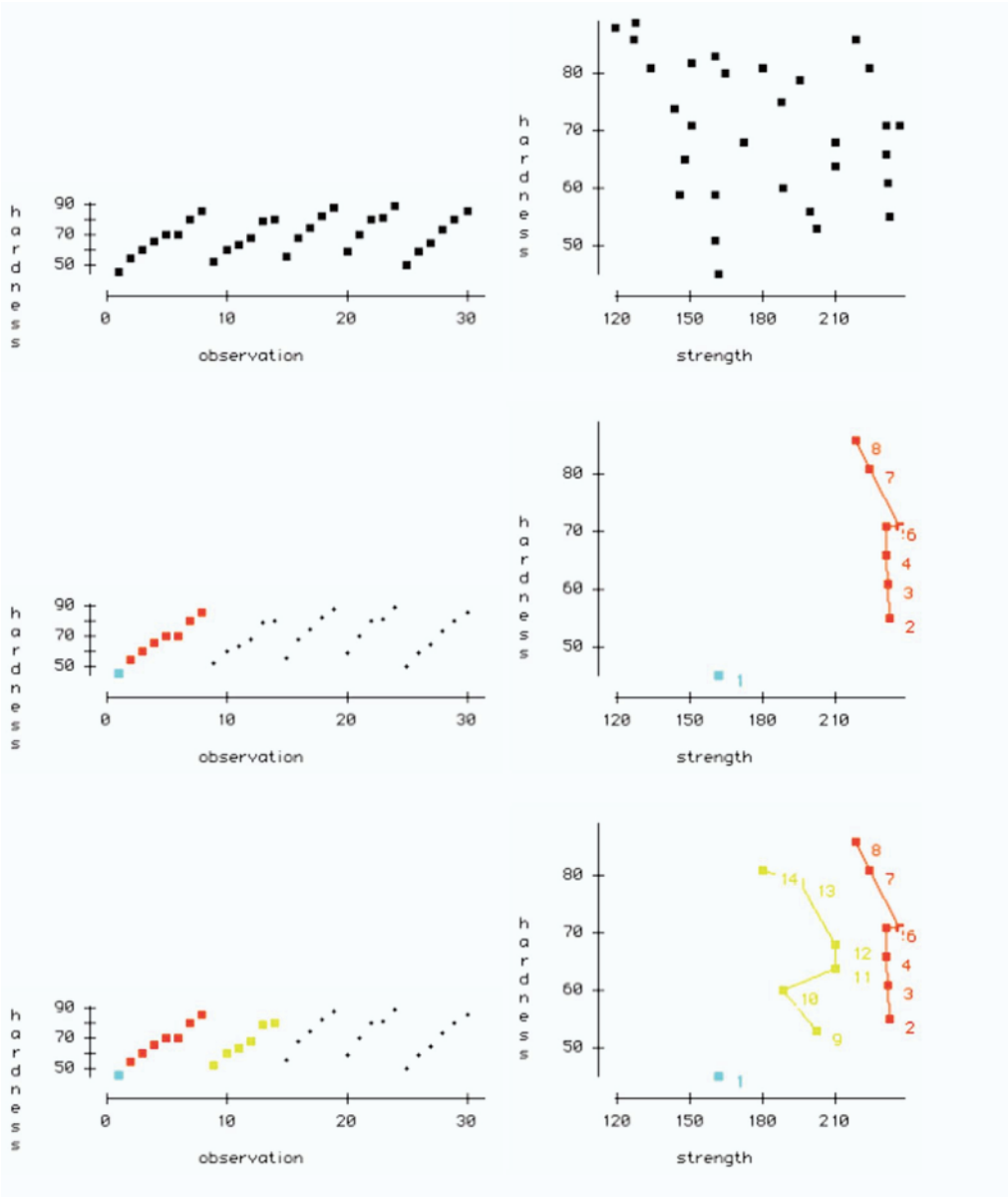


Figure 1. Sequence of dynamic plots brushing through runs of hardness with the selected points displayed in the plot of hardness on strength, labeled by observation number and the points in a run joined by a line.

Murray Jorgensen (pers. comm.) has suggested that a “simulated manual plot,” an animated display where the data are plotted as you watch sequentially in data order would be useful in statistical packages. Brushing with a tall or thin brush comes close to this, highlighting the selected observations using a “sticky” selection mode so that the brushed points stay selected. Slicing (in Data Desk) is a more convenient way of selecting points horizontally or vertically as you brush, leaving the points selected as you brush over them, without having to reshape the brush or change selection modes. And even in Excel (as

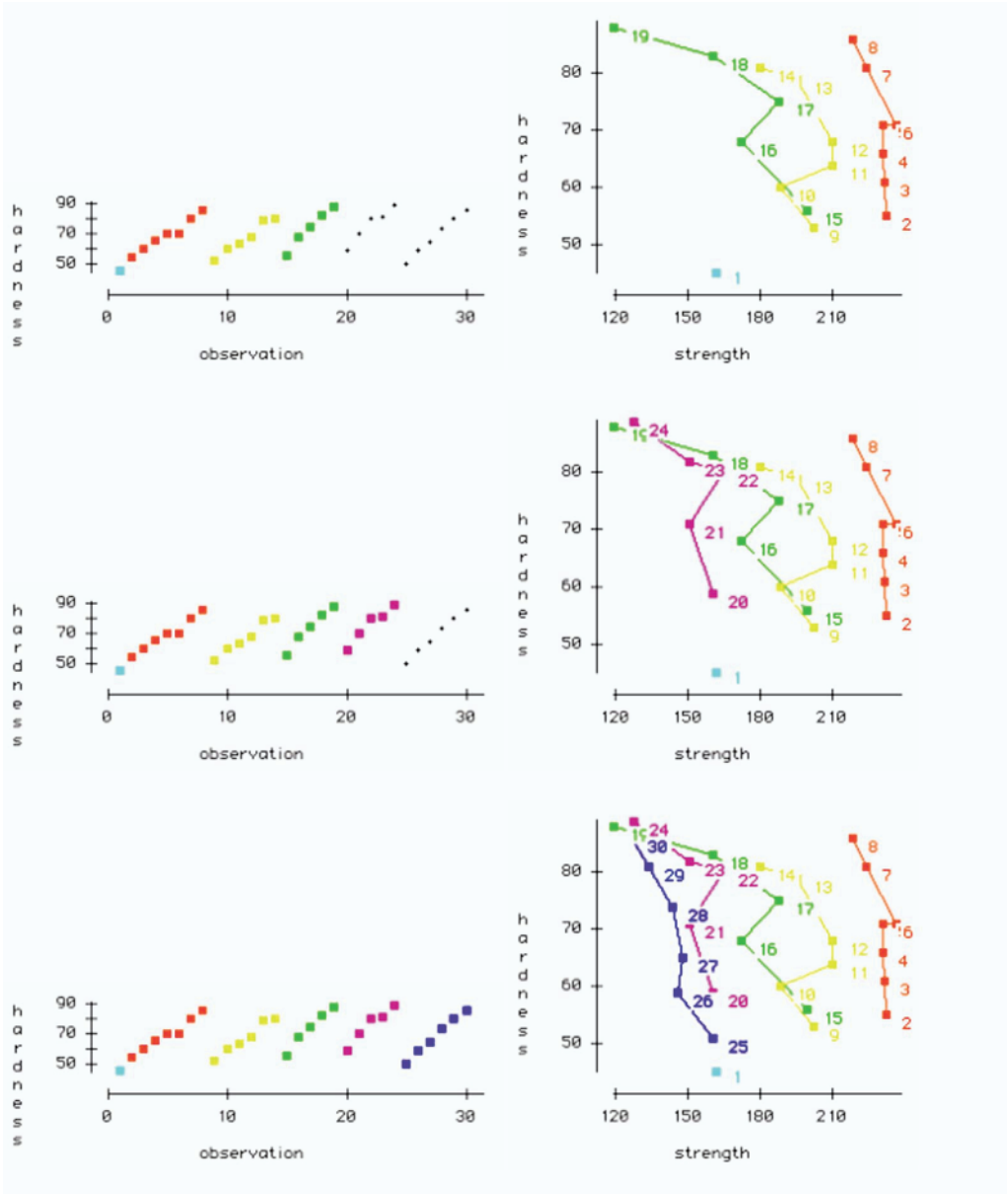


Figure 1. (continued)

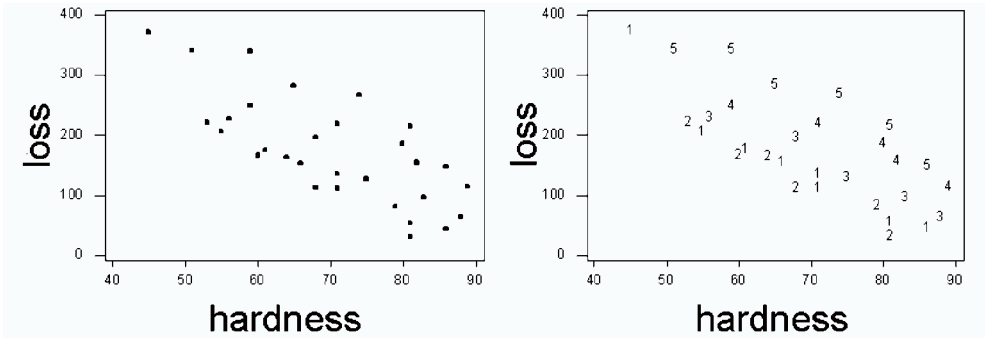


Figure 2. Plots from Preece. (a) Loss on hardness, (b) with Preece's group number the plotting symbol.

in abrasion.xls) you can simulate manual plotting using a dynamic range whose length is controlled by a scroll bar or simply select a series and then step through it in data order with the right and left arrows.

Preece (1987) in a section on data scrutiny comments that

The computer can of course be used to help with data scrutiny. But even then there are dangers. The computer might, for example, give us a scatter diagram such as that in [Figure 2a], which is for a well-known set of data given by Davies . . . However, if we ourselves had plotted the points on graph-paper (do you remember graph-paper, dear reader?) and plotted them in the order in which they appear in the table, we would have found what is illustrated in [Figure 2b], where the points denoted 1's were plotted first, then the 2's, then the 3's, and so on. Proper interpretation of the data is impossible from [Figure 2a]; the *order* of the data clearly needs to be taken into account, although we have insufficient information to say *how* it should be allowed for.

Preece's grouping on order is likely to be based on the hardness series alone and so differs from that discovered here, which looks at hardness and strength, and separates observation 1 from the group of observations 2 to 8.

3.2 STRUCTURE CONFIRMED

What is the actual structure? Davies (1947) refers to the article by Buist and Davies (1946)—the same Davies, he was the statistician on the project—which notes that “the details of the mixes used . . . have been given previously,” referring to an article by Buist and Mottram to be published in the same journal. That article turns out to be Buist and Mottram (1946). The data are listed there (and reproduced here in Table 1) along with *amounts* and *types* of the compounding material, carbon black. Our suggested grouping, discovered from the interactive graphics, of five types is confirmed and sample 1 has a blank added. The points are labeled dynamically by the amount (vols) of carbon black in the plot of hardness on strength in Figure 3, and also hardness is plotted against this amount. Lines join the points in increasing amount for each type of carbon black which are shown by different colors.

The samples within each “type” have a similar strength, but cover a wide range of hardness. A plot of loss on hardness with lines joining the points within a type (Figure 4)

Table 1. Abrasion Loss Data from Buist and Mottram (1946, Table 9)

<i>Observation</i>	<i>Loss</i>	<i>Hard</i>	<i>Strength</i>	<i>Vols</i>	<i>Type</i>	<i>Parts</i>
1	372	45	162	0	blank	0
2	206	55	233	10	Kosmos T	15
3	175	61	232	15	Kosmos T	22.5
4	154	66	231	20	Kosmos T	30
5	136	71	231	25	Kosmos T	37.4
6	112	71	237	30	Kosmos T	45
7	55	81	224	40	Kosmos T	60
8	45	86	219	50	Kosmos T	74.8
9	221	53	203	10	Kosmos 40	15
10	166	60	189	15	Kosmos 40	22.5
11	164	64	210	20	Kosmos 40	30
12	113	68	210	30	Kosmos 40	45
13	82	79	196	40	Kosmos 40	60
14	32	81	180	50	Kosmos 40	74.8
15	228	56	200	15	Kosmos 20	22.3
16	196	68	173	30	Kosmos 20	44.4
17	128	75	188	45	Kosmos 20	66.8
18	97	83	161	60	Kosmos 20	88.9
19	64	88	119	75	Kosmos 20	111
20	249	59	161	15	Magecol	22.3
21	219	71	151	30	Magecol	44.4
22	186	80	165	45	Magecol	66.8
23	155	82	151	60	Magecol	88.9
24	114	89	128	75	Magecol	111
25	341	51	161	20	Thermax	29.7
26	340	59	146	40	Thermax	59.5
27	283	65	148	60	Thermax	88.9
28	267	74	144	80	Thermax	118.9
29	215	81	134	100	Thermax	138.4
30	148	86	127	120	Thermax	177.8

Note: The data, with dynamic graphics, are in abrasion.dsk and abrasion.xls at <http://www.agresearch.co.nz/downloads/graphics/>.

gives a simple static visualization of the data.

This dataset is reminiscent of Prater's crude oil dataset for which Daniel and Wood (1980, chap. 8) reveal a nested structure. The structure for this abrasion data has been largely overlooked for nearly 60 years and, as we saw, is easily revealed by interactive graphics. This is like the wonderful detective story on the barley data by Cleveland (1993), heralding the success of trellis graphics, which unearthed a gem that had also waited to be discovered for 60 years. Cleveland (1993, p. 339) concluded, "The barley example speaks forcefully to the importance of visualization as part of the process of data analysis. Data analysis without visualization, even with methods as brilliant as Fisher's, risks missing important happenings in data, and risks missing ways of simplifying the structure of data, and risks the use of probabilistic methods whose assumptions are not supported by the data. . . . Visualization is a necessary part of data analysis. Tools matter."

So even after nearly 60 years we too have gained insight into old data with simple interactive graphics. However, Box (1999, p. 67) was not impressed by massaging "dead datasets":

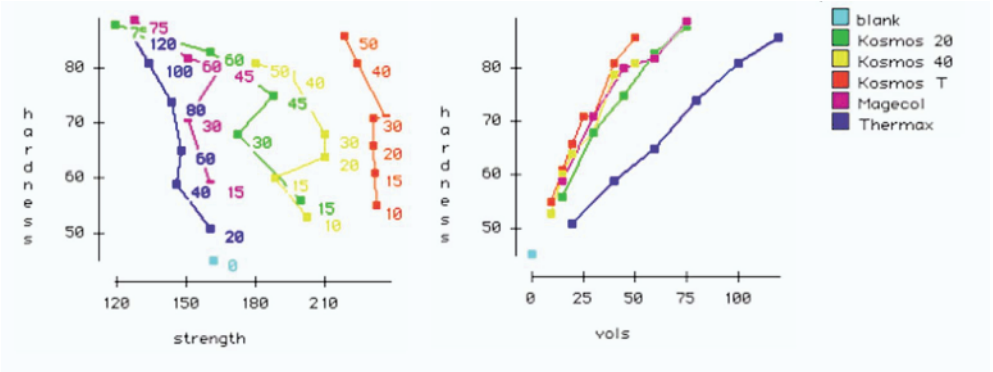


Figure 3. Amounts of the type of carbon black (vols) is added as a label for each point and the type of carbon black are shown by different colors.

When an experimental design is run the most relevant question is “So now what do we do?” I have been saying this for a very long time (see, e.g., Box (1957)). I had hoped that we had seen the end of the obscene tribal habit practiced by statisticians of continually exhuming and massaging dead “data sets” after their purpose in life has long since been forgotten and there was no possibility of doing anything useful as a result of this treatment.

It is interesting that Box came on as a co-author of the 3rd edition of Davies et al. (1957). A referee insightfully commented that:

A reasonable assumption is that the data are ordered in some rational way by the experimenter when the data are presented to the statistician. It is this phenomenon that enabled the “anomaly” to be identified. What if the experimenter had listed the blank first and then Thermax compound results next? In this case no anomaly would be found! Interactive graphics is not a substitute for a consulting statistician not acquiring a thorough understanding of how the experiment was implemented from the researcher. Interacting with the data is a useful way for the statistical consultant to generate questions to ask the researcher. From a Box perspective, this may be a case of current statisticians using interactive graphics to understand what experimental information Davies chose to ignore when the example was selected for use in a book for teaching purposes. Needless to say, it is still good detective work and instructive for students.

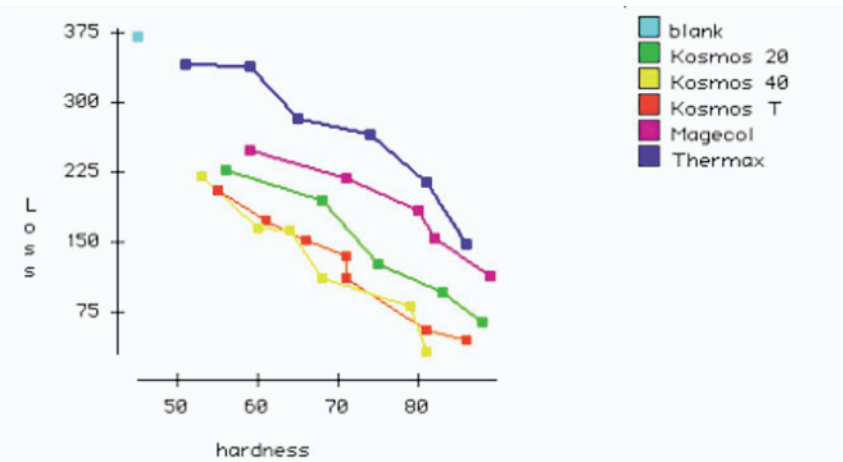


Figure 4. Loss on hardness with lines joining the points within a type, exhibiting the data structure.

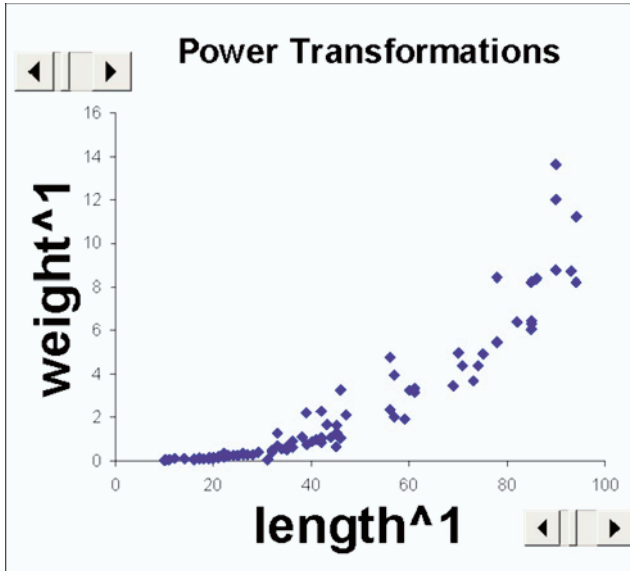


Figure 5. Weight on length, with scroll bars dynamically controlling the powers for each axis.

4. DYNAMIC GRAPHICS, TRANSFORMATIONS, AND OUTLIERS

Dynamic graphics are also useful in exploring data and in determining appropriate transformations of data and checking on possible outliers. Many statistical packages provide tools for doing this but you can even do this easily in Excel, which we will use in this section, using another small dataset (“slug work.xls”) on slugs available from <http://www.agresearch.co.nz/downloads/graphics/>.

Over a period of four years 100 slugs of the species *Limax maximus* encountered in the field were collected by two Ruakura researchers, Barker and McGhie (1984). Included was a regular sampling of a population living under a macrocarpa hedge in Hamilton. The slugs were preserved in a relaxed state and subsequently their weight (gm) and length (mm) recorded and shown in Figure 5. Growth occurs throughout the year and there is a wide range of body size in any given population. There is a seasonal reproductive cycle and slugs live for several years. Greatest numbers of juvenile slugs occur from May to September.

A simple model is $\text{weight} = \text{volume} \cdot \text{density} \propto \text{volume}$, assuming constant density, and so $\text{weight} \propto \text{length} \cdot \text{width} \cdot \text{depth}$. But we only have length, and assuming growth rate is similar in all dimensions $\text{weight} \propto \text{length}^3$ is a reasonable model. That is shown in Figure 6. One of the early uses of dynamic graphics was plotting dynamically power transformed data (Fowlkes 1971). It is widely available now in statistical packages and is even easy to do in Excel, as is shown here, with scroll bars dynamically controlling the powers for each axis. We would expect transformed scales $\text{weight}^{1/3} \propto \text{length}$ (Figure 7) or $\log(\text{weight}) \propto 3\log(\text{length})$ (Figure 8) to be better, and indeed they are. An outlier becomes evident in

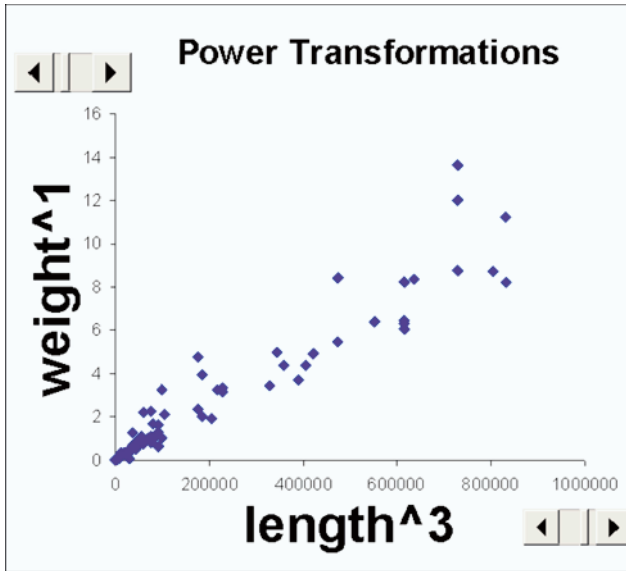


Figure 6. Weight on $length^3$, with scroll bars dynamically controlling the powers for each axis.

Figure 7 and even more so with log transformed axes in Figure 8. Identifying this outlier is easy by hovering over the point and a pop up label identifies the point as slug 90 with length 31 and weight .03. In Excel you can click on the point and move it. But beware: the data change too! However, this is great for illustrating leverage and outliers in regression as the regression line changes as you move a point. Moving the outlier (slug 90) you can see that maybe its weight should have been .3 instead of .03 (Figure 9) or perhaps it was a dyslexic slug and its length should have been 13 instead of 31 (Figure 10). Either of these is plausible, but we can no longer check, so we would omit this point from the analysis as it is clearly implausible. Doing this dynamically you get instant feedback on the effect of these changes.

Brushing or further regression diagnostics (Figure 11, using Minitab) with slug 90 removed, reveal clusters of residuals.

Apparently 2 to 10 slugs were collected on a particular day and these were stored in a dish for that collection day. Later the dishes were taken, in no particular order, and the preserved slugs were measured. Perhaps there are seasonal and/or site differences in the relationship between weight and length. And maybe larger slugs are more likely to be picked up and measured earlier from each dish (Figure 12). But the actual structure of the data is no longer available.

When my colleague, Neil Cox, analyzed these data in the early 1980s the data were plotted on graph paper. Barker and McGhie (1984) plotted the cubed root of weight on length with no comment on the outlier. Since then Neil and I have used these data to illustrate transformations and outliers. Dynamic graphics make it so much easier to explore these data. Doug Stirling has used these data with a very nice implementation of transformations with

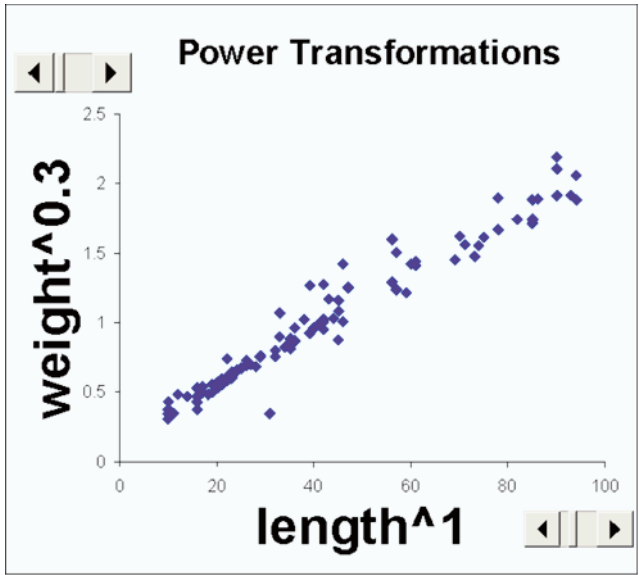


Figure 7. $\text{Weight}^{-0.3}$ on length , with scroll bars dynamically controlling the powers for each axis.

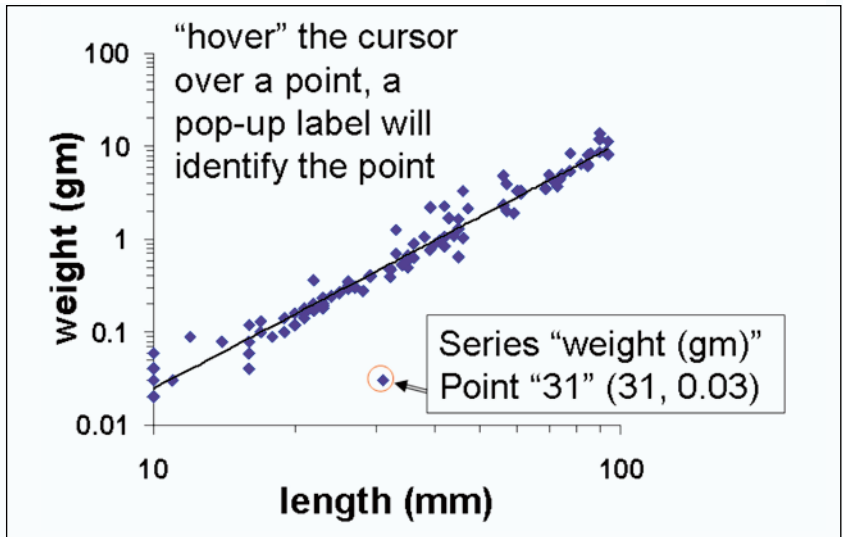


Figure 8. Identifying the outlier.

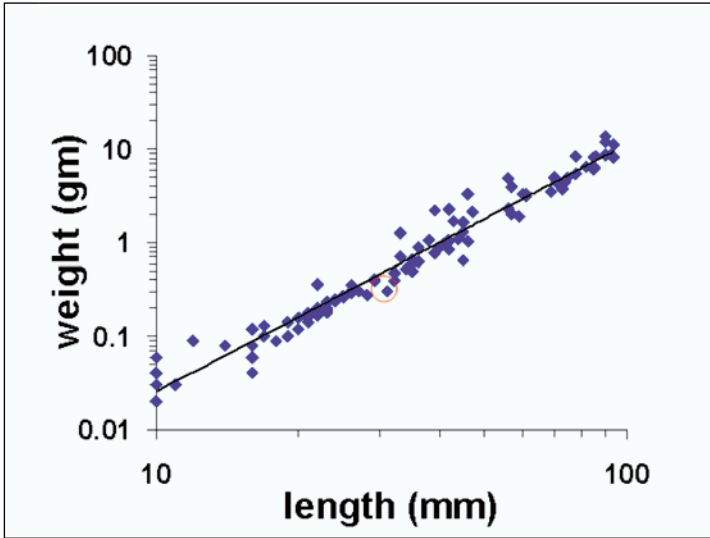


Figure 9. The result of moving the outlier weight from .03 to .3.

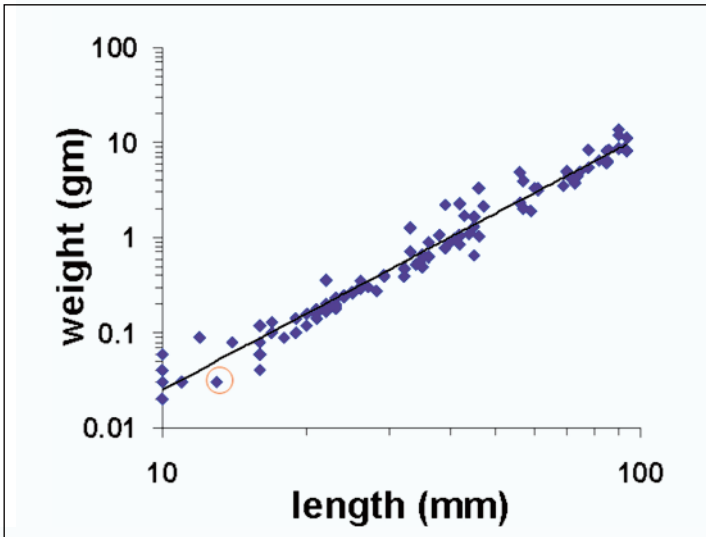


Figure 10. The result of moving the outlier length from 31 to 13.

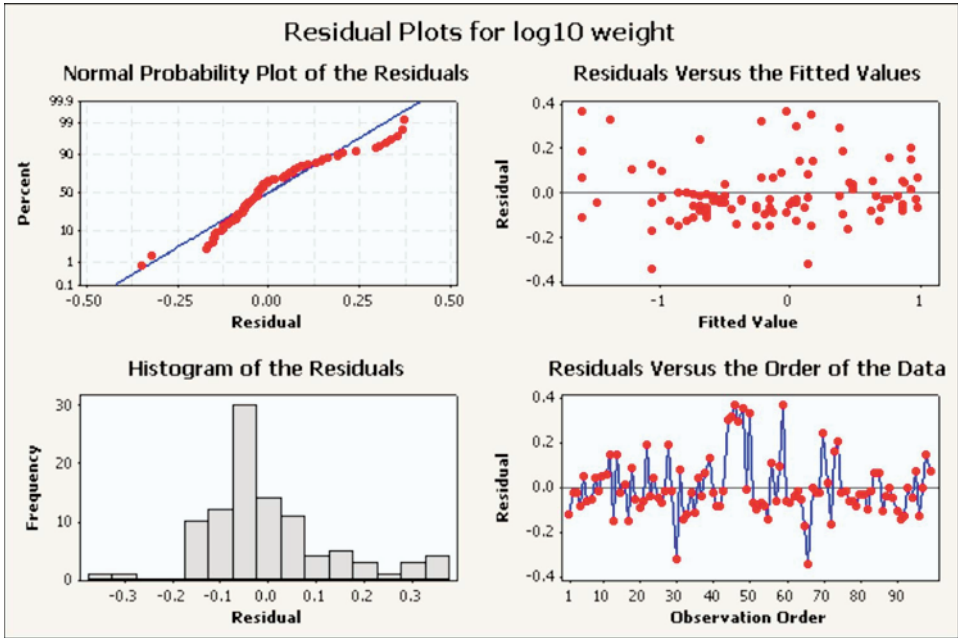


Figure 11. Regression diagnostics, with slug 90 removed, reveal clusters of residuals.

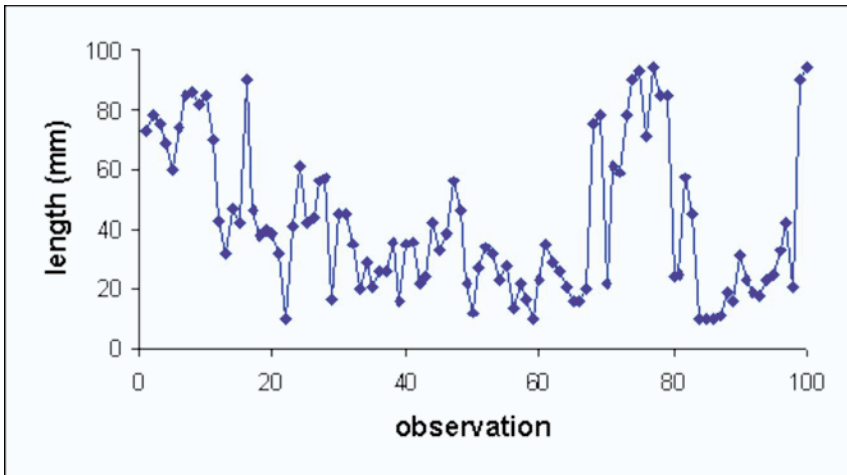


Figure 12. Length in data order.

dynamic graphics in a module in Computer Assisted Statistics Teaching (CAST), available at cast.massey.ac.nz at no charge for use in teaching. Statistics packages often have nicer implementations of power transformations than what is set up and illustrated here in Excel. The widespread use of Excel means that even without specialized software some dynamics graphics are easy to set up. I have even used this slug example and other examples, set up in Excel, in talks to high school teachers. No expert knowledge is required to set up the slug example in Excel. All you need to do is select a scroll bar from the “Forms” toolbar to dynamically control the power for each axis.

5. INTERACTIVE GRAPHICS FOR MICROARRAY DATA

Microarray data can have a large number of observations, particularly when data for the slides are stacked. It is interesting to explore the use of interactive graphics on datasets like this. We will consider only the raw image data before it is normalized, and give examples of some interesting aspects that are easily revealed using simple interactive graphics.

The use of interactive graphics to explore microarray image data is illustrated here using Data Desk. Similar exploration could be done in other interactive graphics environments. Cook et al. (in press) describe visual methods for analyzing microarray data, using GGobi. They give a worked example with data from an experiment using Affymetrix GeneChips. They stress that “Interaction on plots is critical! Regardless of the actual display that is used, it is important in exploratory data analysis to be able to interact with the plot and link the interaction to changes in simultaneously visible displays.”

Introductions to microarray technologies and analysis are Nguyen, Bulak, Naisyin, and Carroll (2002) and Craig, Black, and Doerge (2003).

Use of interactive graphics to explore microarray image data is illustrated here with data from an experiment from the Dairy Biotechnology Group at AgResearch Ruakura, investigating gene expression in mammary glands with mastitis. Each slide compares a control gland and an infected gland from the same cow. Two fluorescent dyes are used, Cy5 (Red) and Cy3 (Green), one for each sample. Comparisons were made for each of six cows, with a dye reversal of each, using 12 cDNA microarrays each with approximately 22,000 spots. The 12 slides are from the same print run (48) and have the same layout. The slides were scanned and the visual images were converted to data files using GenePix 3.0 (Axon Instruments 2002). The composite image for slide 48-80 is shown in Figure 13.

Interactive graphics is implemented in microarray image analysis packages like GenePix to give the user some control over the spot detection, alignment, and image analysis on a single slide. We are interested in exploring the image data over all the slides in an experiment and not just a single slide and so set this up in Data Desk. We work with the image data stacked by slide giving approximately 270,000 rows for this experiment with 12 slides.

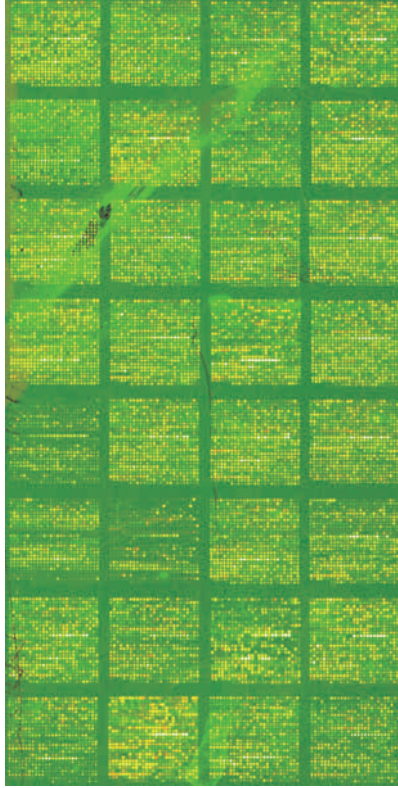


Figure 13. The composite image for slide 48-80.

5.1 VIRTUAL ARRAY

The slides in the same print run all have the same spatial layout, so we set up a virtual array on row and column number to select spots rather than actual positions of the spots. We use interactive graphics with linking and hot selection to explore the alignment of the pins (8×4 , metarow \times metacol) and spots (26×27 , row \times col, for each pin) across the slides. The virtual row is $-(30 * (\text{metarow} - 1) + \text{row})$ (the negative sign is so that the rows go down to match the orientation of the image), and the virtual column is $30 * (\text{metacol} - 1) + \text{col}$.

For example, selecting (with the mouse) the top left spot in the virtual array (Figure 14, left plot) shows the actual positions of that spot on the 12 slides (Figure 14, right plot). Clearly the spots do not line up across slides on actual x, y positions from GenePix. And even if the actual positions were adjusted by an offset for each slide, these adjusted positions of corresponding spots would still not coincide exactly for all spots.

For example, selecting slide 84 highlights the virtual array and actual array positions. Even at this resolution it appears (Figure 15) that the actual position of spots for pin row 1 column 2 are out of alignment with the other pins in that row for slide 84. And (less easy to see at this resolution) but consistent over slides, the actual positions of spots in pin row 5 column 4 appear to be shorter than other pins.

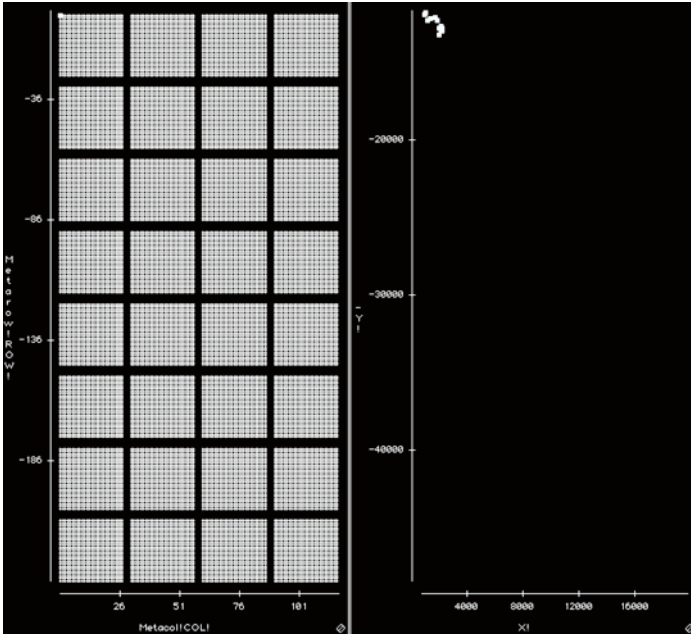


Figure 14. Selecting the top left spot in the virtual array shows the actual positions of that spot on the 12 slides.

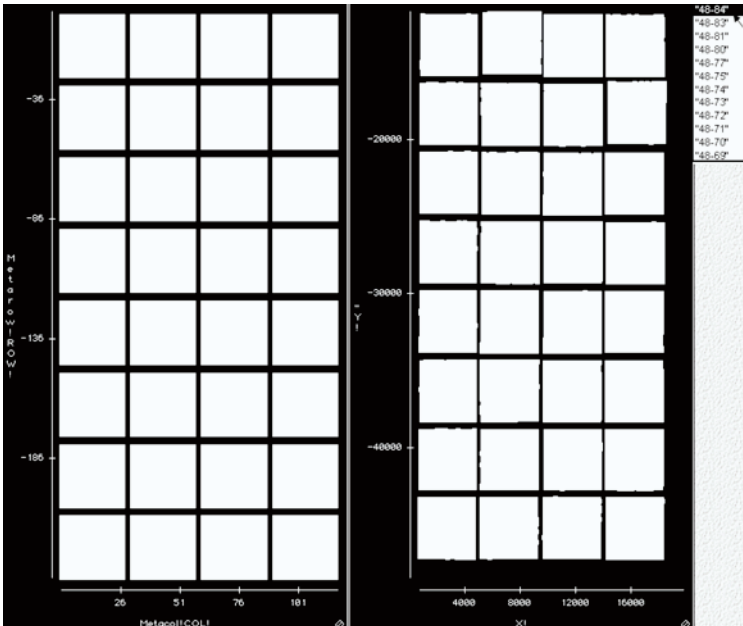


Figure 15. Selecting slide 84 highlights the spots for slide 84 in the virtual array and their actual array positions.

5.2 ZOOMING IN ON ACTUAL POSITIONS

Zooming in on the actual positions of spots for pin row 5 column 4 for each slide (Figure 16 shows slide 48-69 alongside its slide image) shows misaligned rows, starting with spot 7 from the left in row 16.

It is easy to check that all the slides have a similar misalignment for pin row 5 column 4 by interactively selecting each slide in turn.

The print run takes three days. It is likely that the pin was removed and cleaned and then returned in a slightly different orientation after printing the spot 8 from the left in row 16, so that the subsequent spots (the print direction is right to left in successive rows moving down) are slightly misaligned up towards the previous row. The rest of row 16 overlaps row 15 and row 17 then overlaps row 16 until spot 7 from the left. Interpretation of these spots will be difficult, as the overlapping spots appear to have mixed, and they should be excluded from analysis.

5.3 FLAGS

GenePix automatically flags a spot “not found” (flag = -50) if “Align Block” fails to find a spot. The last six positions for each pin have not been printed and are flagged “absent” (flag = -75). Spots the user flags as “bad” (flag = -100) are to be omitted. Unflagged spots have flag = 0. Spots may be dynamically colored by flag, as in Figure 16. The overlapping spots in Figure 16 can be selected and their flag set to -100 to exclude them from analysis.

5.4 COLOR

The foreground (spot) red and green intensities are denoted by R_f and G_f , respectively, for each spot and the background intensities by R_b and G_b . The background-corrected intensities are $R = R_f - R_b$ and $G = G_f - G_b$. The log-differential expression ratio is

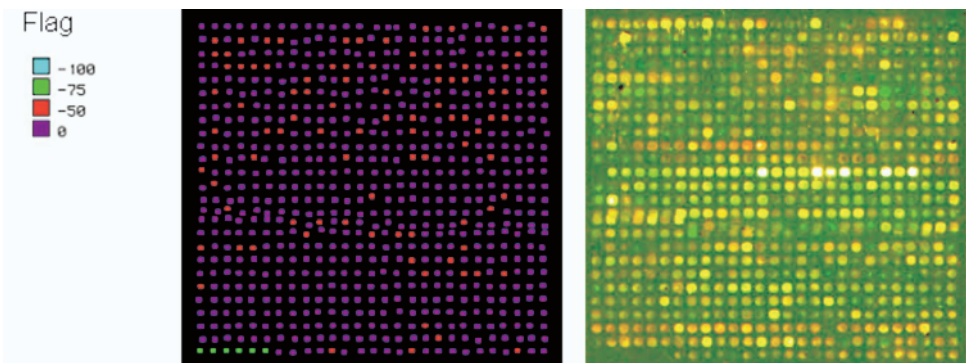


Figure 16. Actual positions of spots, colored by flag, for pin row 5 column 4 for slide 48-69 alongside its slide image.

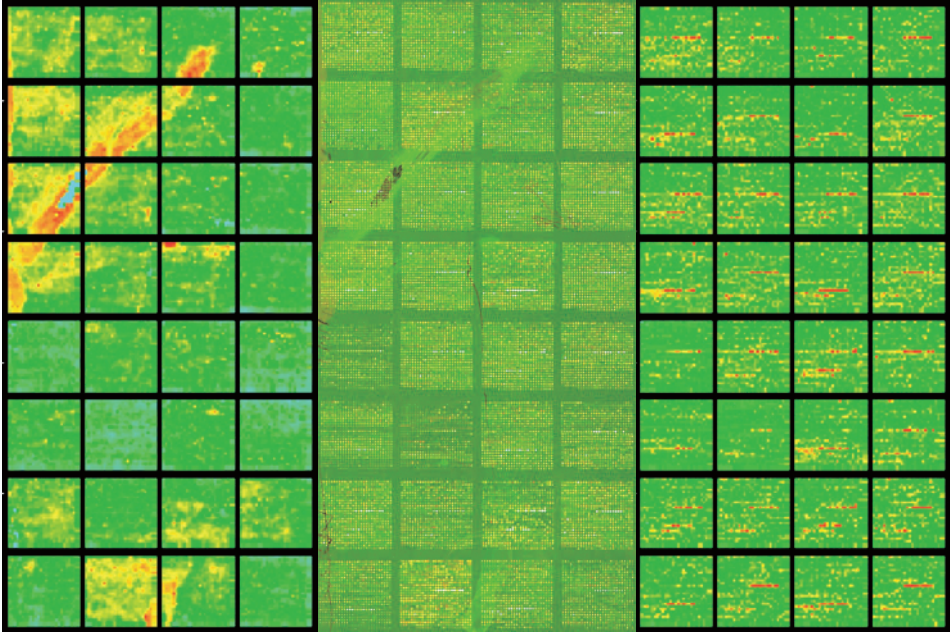


Figure 17. Slide 48-80 (center) with points colored hot to cold by background intensity (left) and foreground intensity (right).

$M = \log_2(R/G)$ for each spot. The log intensity, $A = \frac{1}{2}\log_2(RG)$, measures the overall brightness of the spot. Plotting M on A gives the well-known MA plot.

It is easy to dynamically select a slide (Figure 17 has slide 48-80) and color the points hot to cold by background intensity (left) and foreground intensity (right). It is simple to dynamically color by other variables. In Data Desk buttons to color by, for example, spot or background red or green intensity, A , M , and flag can be set up.

5.5 A YEAST CONTAMINATION

Look at the bright green area (Figure 18) at the top left of pin row 4 column 3 for slide 48-80. It turned out to be a yeast contamination, originating in the well in the 384 well plate in the library for the EST spotted five from the left in the first row. It raises interesting questions of carryover in the printing. The pins are washed between changing spots and the washing is supposed to remove 99% of the previous sample. The print direction is right to left in successive rows moving down as illustrated with the arrows (Figure 19).

Selecting the five left-most positions in the first row in the virtual array for this pin shows a similar pattern of intensities on all 12 slides, as shown in Figure 20. These five spots were (nearly always) flagged bad by Kara Swanson as she scanned the images. It is easy to select these five spots for all slides dynamically and to set their flags bad (-100). The spots are colored by flag after this had been done. Spots on the right of the next row, showing yeast contamination, should probably also be flagged bad.

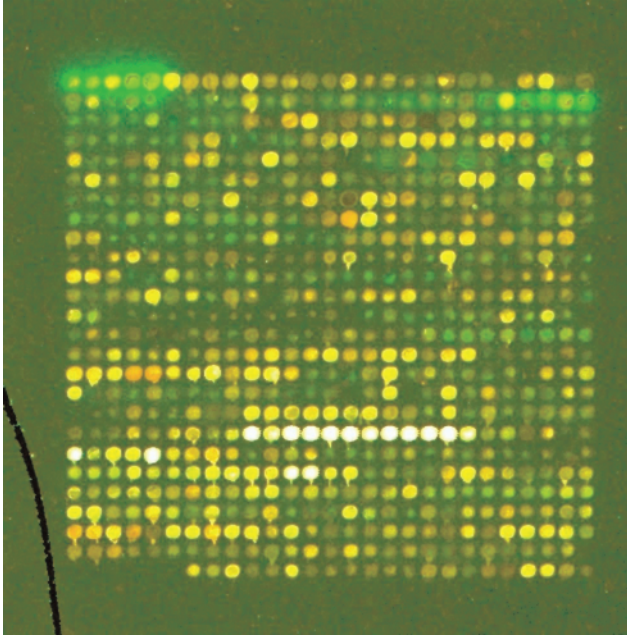


Figure 18. Pin row 4 column 3 for slide 48-80.

To further explore what appears to be carryover of the yeast contamination, we dynamically select the three top rows for this pin, and plot background and foreground log greens in print order within slide (calculated as $\text{row} * 27 - \text{col} + 1$). Stepping through each slide shows a similar pattern to that for slide 48-80 which is displayed in Figure 21. The intensities appear to decline, with carryover, after the initial high spot, and continue to decline from the right in the second row. And the background intensity rises at the left of the second row. So we should also flag these extra spots with yeast contamination as bad.

5.6 CHECKING BY SLIDE

Figure 22 is an annotated screen shot, to label some of the linked displays, when selecting slide 48-80, colored by intensity. Hot selection with automatic update has been turned on to show just the selected points for all the graphs, except the scatterplot of all



Figure 19. The print direction is right to left in successive rows moving down as illustrated with the arrows.

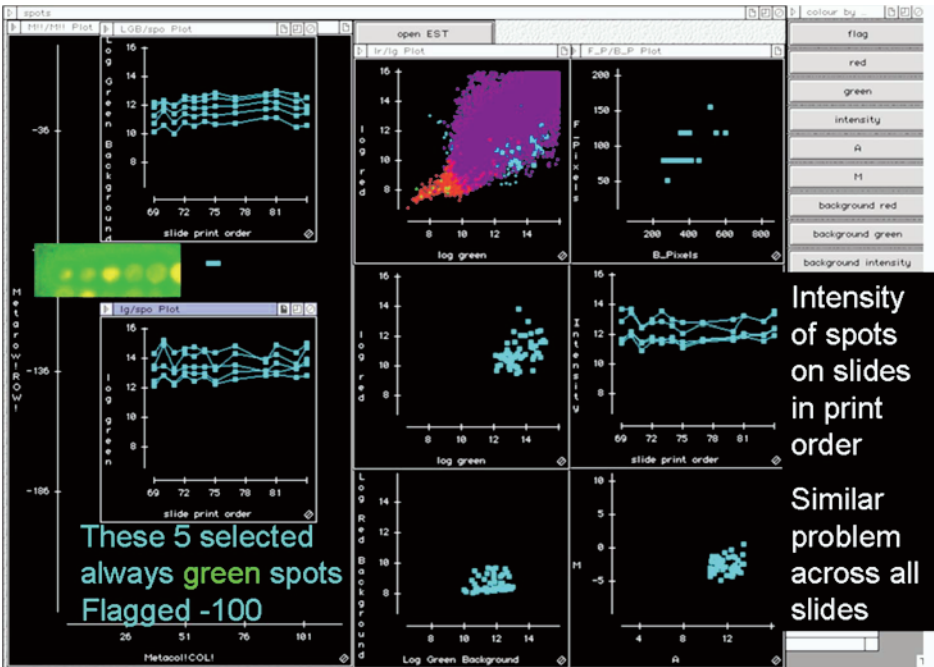


Figure 20. Selecting the five left-most positions in the first row in the virtual array for this pin shows a similar pattern of intensities on all 12 slides.

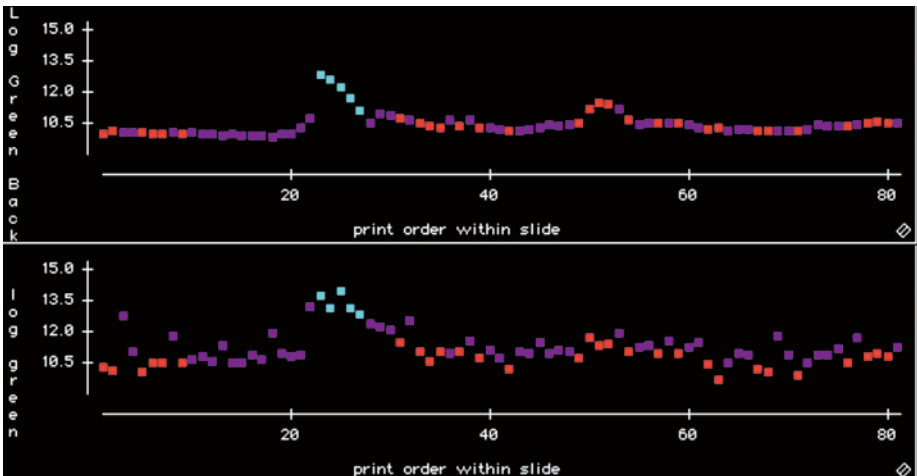


Figure 21. Background and foreground log greens in print order for the three top rows for this pin in slide 48-80.

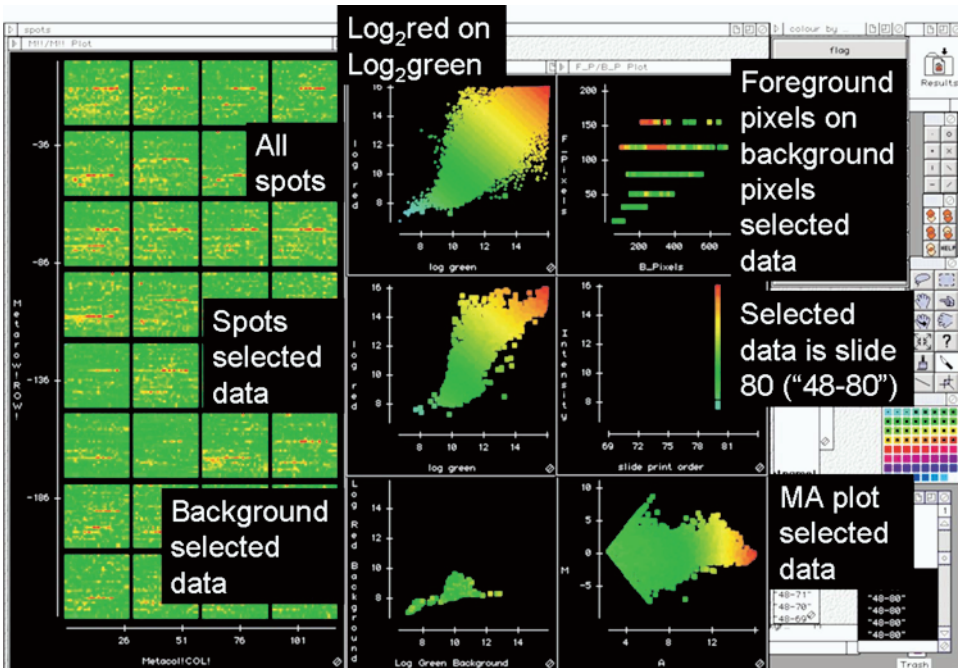


Figure 22. Annotated screen shot labeling some of the linked displays, colored by intensity. The virtual array of row on column number is the plot on the left. The middle column of graphs show $\log_2 R_f$ on $\log_2 G_f$ for all the data in the top graph and for the selected data immediately below. The bottom graph is $\log_2 R_b$ on $\log_2 G_b$ for the selected data. The column of graphs on the right show for the selected data the number of pixels in the foreground on the number of pixels in background, the intensity on slide and finally the MA plot.

spots of $\log_2 \text{red}$ on $\log_2 \text{green}$ (specifically, $\log_2 R_f$ on $\log_2 G_f$). The plot scales have been frozen for all the data so the scales do not rescale for selected points. Screen shots like this, for other selected points, will be displayed subsequently without annotation.

Evident are the strong row effects on intensity. These arise mainly because ESTs from the same 384 well plate from the library are printed in sequence, and in the same positions for all the 32 pins. (From each 384 well plate the 32 pins print 12 spots each. These slides have 27 spots in each row for each pin and so the sets of 12 spots will sometimes straddle 2 rows.)

The sets of high (hot, red in this plot) intensity spots in row 12 are from the same 384 well plate. Selecting this set of 12 is easily done dynamically from a within-pin positions plot of $-\text{row}$ (the negative sign is to have to rows going down) on column within pin, colored by spot intensity. This set of 12 is shown for slide 48-80 in Figure 23.

It is sometimes convenient to dynamically select from the virtual pin array and/or the virtual within pin array.

5.7 CARRYOVER TO WATER?

There are 219 spots that are water in this print run. Some are named water or $\text{H}_2\text{O } n$ or $\text{H}_2\text{O } n$, where n is some number. So we will check whether there is any evidence of

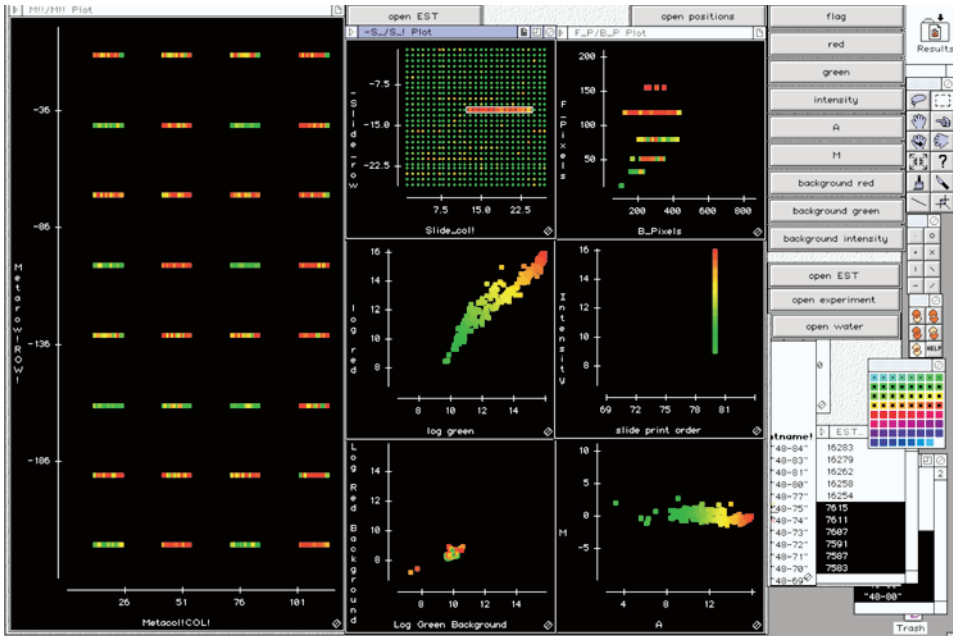


Figure 23. The set of 12 high intensity spots in row 12 for slide 48-80 selected from a virtual within pin array displayed in the top center panel.

carryover from the previous spot to a water spot in the printing. To do this, color by intensity and then select the 219 water spots (by finding water, H₂O or H₂O in the EST variable) and change their color to white. This shows the position of the water spots and the intensity of the other spots. Profiles of the log intensities of the waters are shown across the slides in Figure 24.

One water, H₂O 87, is consistently high on all slides. Zoom in on pin row 1 and column 3 and select H₂O 87 (Figure 25).

Then select the few high-intensity spots that are printed before H₂O 87 shown in Figure 26.

This high water, H₂O 87, is printed immediately after a series of high-intensity spots (the printing is right to left). The immediately preceding two spots are saturated on all the slides. We use interactive graphics on the virtual array to select the waters and the spots that are printed before the waters shown in Figure 27 for pin row 1 column 3.

The GenePix file is sorted on scan order which is from left to right within rows within pin, rather than the print order, which is from right to left. Sorting the data into print order (by sorting on $\text{slide} * 10^6 + \text{pin} * 10^4 + \text{row} * 10^2 - \text{col}$) makes it much easier to set up lagged variables of the previous spot alongside the spot. The first spot printed by each pin, of course, has no previous spot, and so the corresponding observation of the lagged variable needs to be deleted. This can be done easily by selecting the first spots dynamically from the within pin virtual array.

Plotting A (the average log background-corrected intensities) for water on A for the

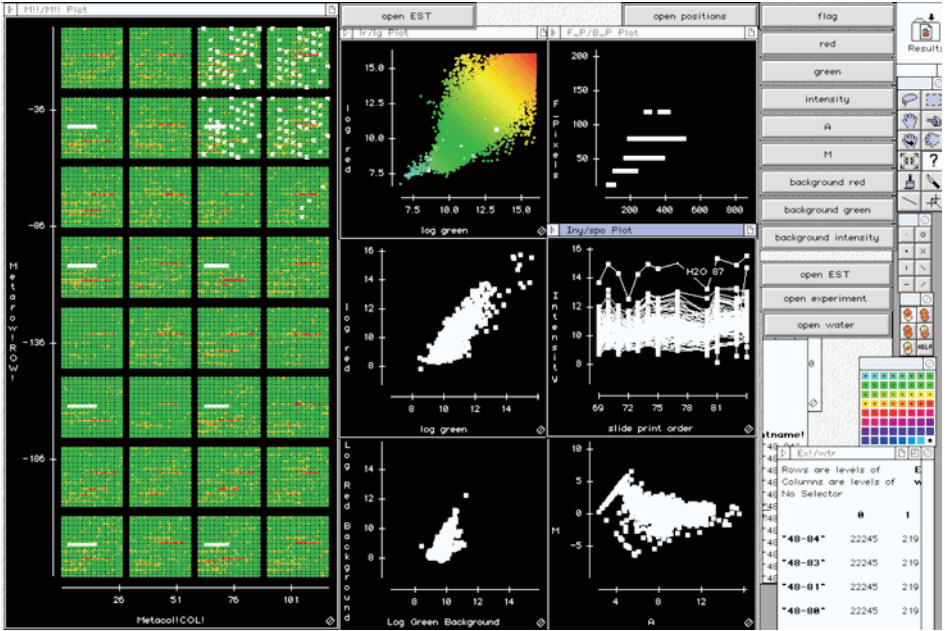


Figure 24. The position of the 219 waters and profiles of the log intensities of the waters across the slides. One water, H₂O 87, is consistently high on all slides.

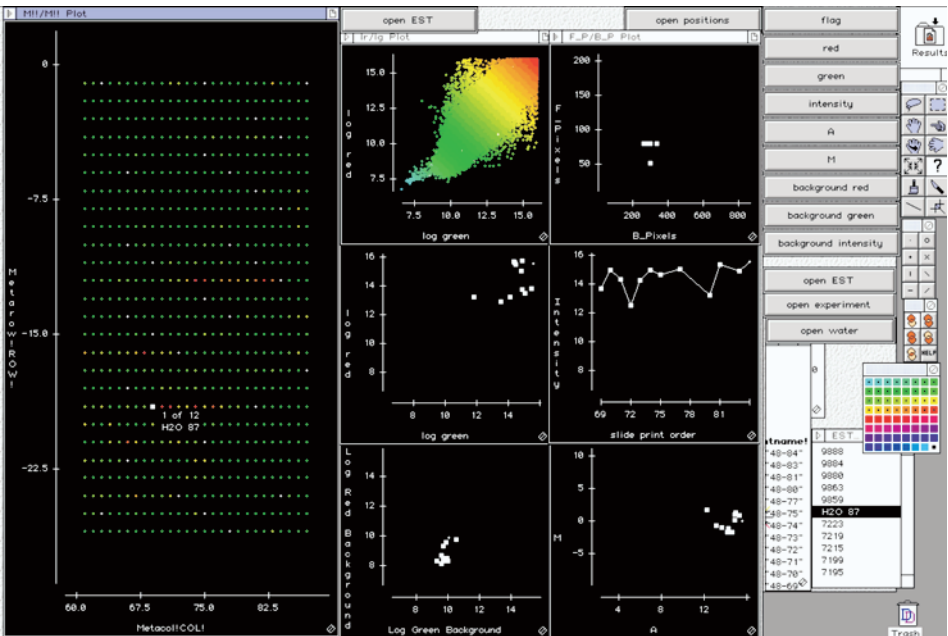


Figure 25. Select H₂O 87 and zoom in on pin row 1 and column 3.

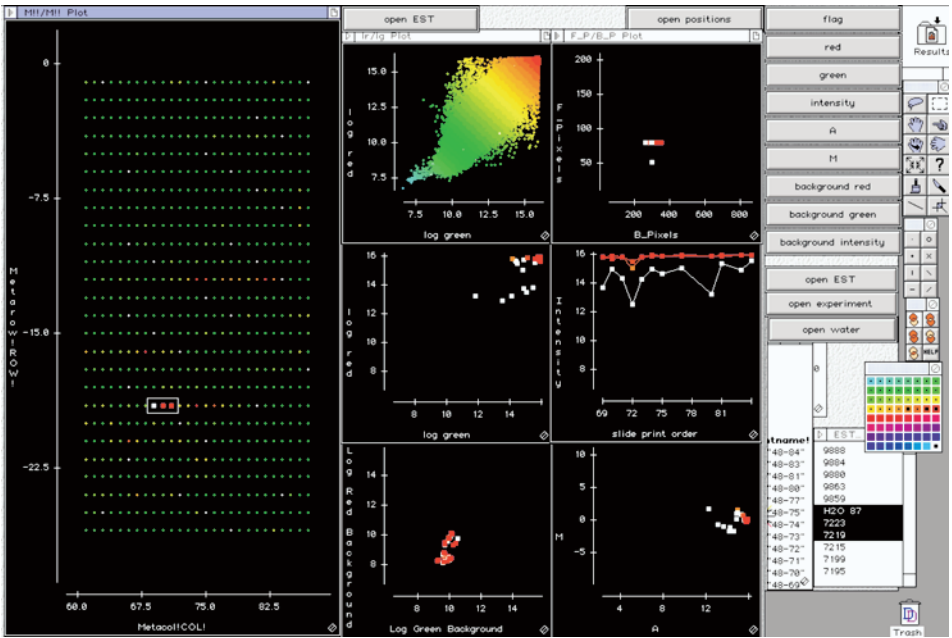


Figure 26. Then also select the few high intensity spots that are printed before H₂O 87.

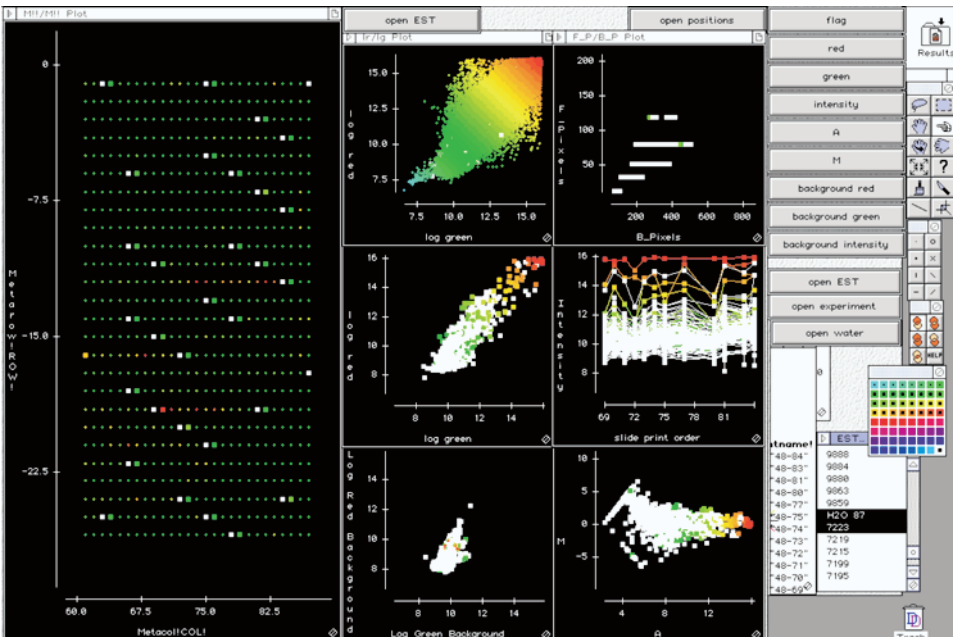


Figure 27. Select the waters and the spots that are printed before the waters for pin row 1 column 3.

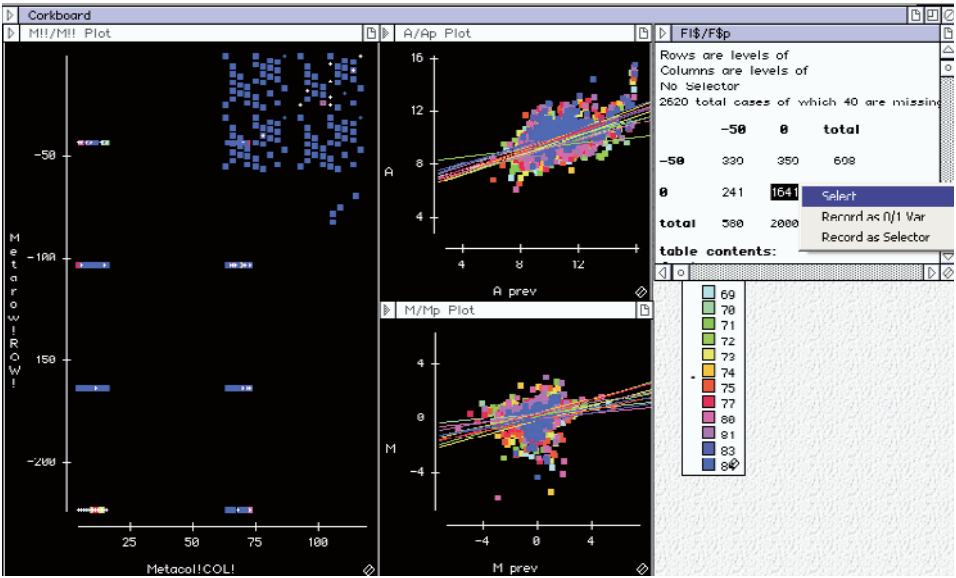


Figure 28. Checking for carryover in A and M. Plotting A for water on A for the previous spot, and M for water against M for the previous spot, both show a positive trend for each slide. Only waters and previous spots that are both unflagged are included.

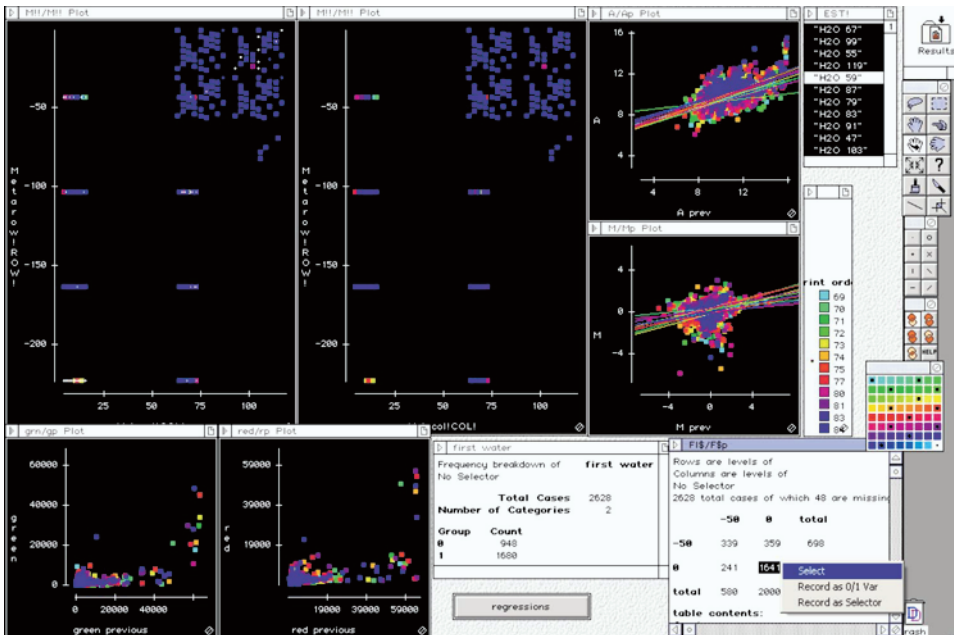


Figure 29. Waters (unflagged) with previous spot (unflagged).

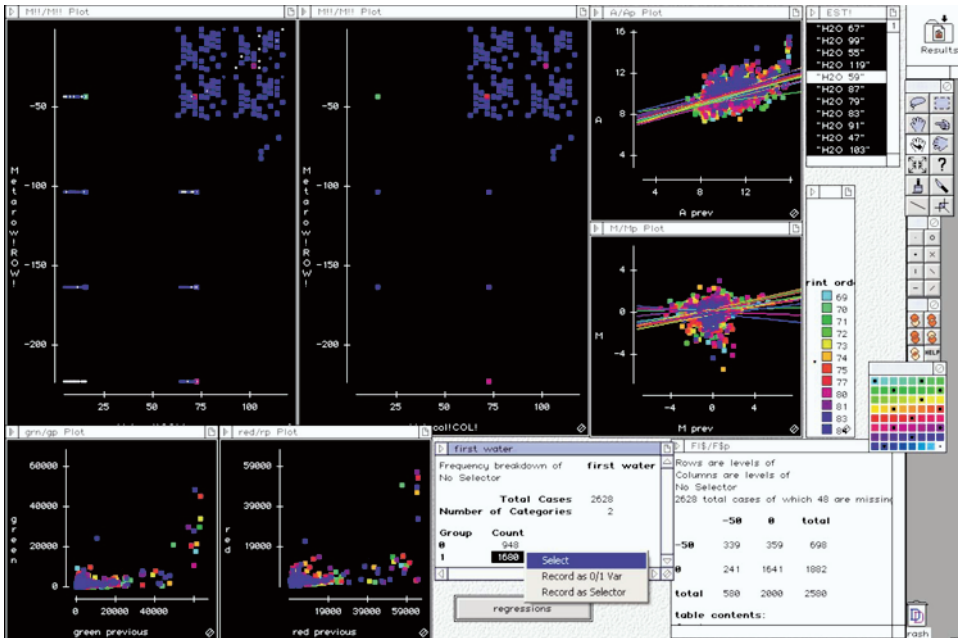


Figure 30. Waters (unflagged) with previous spot something other than water (unflagged).

previous spot, and M (the log-differential expression ratio) for water against M for the previous spot, both show a positive trend for each slide. In Figure 28, only waters and previous spots that are both unflagged are included. Points are colored by slide and regression lines for each slide have been added dynamically.

It is then easy to dynamically select the unflagged data in turn for each slide and check out the regression for each slide, which can be dynamically displayed when hot selection and automatic update are turned on. So there is evidence of carryover to spots that are just water in both A and M from the previous printed spot.

In Figure 29, as in Figure 28, only waters and previous spots that are both unflagged are included. Points are colored by slide. Restricting this further to when the previous spot is something other than water is shown in Figure 30. These alternate views are easily generated dynamically as the tables of flag status of the spot to the previous spot and a table of whether the previous spot is water link to the graphs and regression analyses. For example, in Figure 29, the 1,641 spots with both the spot and its previous unflagged data are selected from the table. In Figure 30 the selection mode is changed to AND after selecting the 1,641 spots and the data are further narrowed by selecting those (1,680 in the table first water) where the previous spot is something other than water, giving plots of the 1,285 spots that satisfy both conditions.

5.8 SETTING IT UP IN DATA DESK

We worked with the image data stacked by slide giving approximately 270,000 rows for this experiment with 12 slides. The resulting file, set up with various calculations (including

row and column for the virtual array, log transforms on Red and Green foreground and background, A and M) graphs and control buttons on corkboards displayed in this section is 314MB—too large to reasonably download. What is perhaps surprising is that the interactive graphics are still fast on data of this size and on files as large as this—even on a 700 Mhz laptop with only 256 MB RAM. I have routinely used this sort of setup on experiments with up to 36 slides and 23,000 spots per slide with around 800,000 observations in the stacked file. It is easy to set up and would take an experienced Data Desk user about half an hour to set up from scratch from the stacked data file, with variable names in the first row. Having set it up once, it is easy to substitute new data for old data and change the calculations for the virtual array using the appropriate number of rows and columns for the new experiment.

6. CONCLUSIONS

Interactive and dynamic graphics with tools of brushing and linking are critical to modern exploratory data analysis. The power of interactive graphics is that it helps you discover unexpected patterns, structure, and relationships among several variables, by touching graphs and watching them change in response to your actions. One of the most useful features of interactive graphics is linking between different displays, whereby, selecting points in one display highlights the parts of other plots that display those points.

Interactive and dynamic graphics have become increasingly available in many statistical and visualization packages, and some are even available, or easy to set up, in general spreadsheets like Excel. The widespread use of Excel means that even without specialized software some interactive and dynamics graphics are available to general users. In AgResearch, our scientists generally maintain their data in Excel and we encourage “good spreadsheet practice” and simple exploratory data analysis in Excel. Filters give dynamic graphics by showing plots of just the data selected by the filter. Pivot tables and charts provide interactive tables and plots summarized by variables that can be dragged and dropped onto or off the table or graph.

A referee commented

All three examples illustrate how dynamic graphics has essentially enabled the consulting statistician to return to an era when they received the data on paper, personally plotted the data, and personally calculated statistics: the process was an integral part of “getting to know” the data. It is that aspect, “getting to know the data,” that I see in the examples presented. Dynamic graphics has overcome the “de-personalized” computerization of data that has dominated statistics over the past 30–40 years. It also goes beyond the historical “personal involvement” era by enabling the statistician to more quickly do additional graphics. The latter might be considered to have “revolutionized statistical practice.” It may be more appropriate to say that it enables good consulting statisticians to do what they always wanted to do—but couldn’t due to time constraints.

ACKNOWLEDGMENTS

I thank the editor for the invitation to present this material at the 2004 International Biometric Conference and to submit this article as an editor’s invited paper. I am grateful for the useful suggestions of two referees,

the editor, and colleagues Neil Cox, Katarina Domijan, Vanessa Cave, and Hadley Wickham which improved this article. I thank the Dairy Biotechnology Group at AgResearch Ruakura for use of the microarray data and David Baird, AgResearch Lincoln, for discussion on analysis of microarray data.

[Received July 2004. Revised October 2004.]

REFERENCES

- Barker, G., and McGhie, R. (1984), "The Biology of Introduced Slugs (*Pulmonata*) in New Zealand 1. Introduction and Notes on *Limax maximus*," *NZ Entomologist*, 8, 106–111.
- Becker, R. A., Cleveland, W. S., and Shyu, M-J. (1996), "The Visual Design and Control of Trellis Display," *Journal of Computational and Graphical Statistics*, 5, 123–155.
- Becker, R. A., Cleveland, W. S., and Wilks, A. R. (1987), "Dynamic Graphics for Data Analysis," *Statistical Science*, 4, 355–395.
- Box G. E. P. (1957), "Integration of Techniques in Process Development," *Transactions of the American Society for Quality Control*, Milwaukee: American Society for Quality Control.
- Box, G. E. P. (1999), "Response," *Journal of Quality Technology*, 31, 67–72.
- Buist, J. M., and Davies, O. L. (1946), "Statistical Evaluation of Variations in Rubber Processes and Correlation in Physical Properties," *Transactions of the Institution of the Rubber Industry*, 22, 68–81.
- Buist, J. M., and Mottram, S. (1946), "Effect of Carbon Blacks on the Physical Properties of Natural Rubber and Neoprene GN," *Transactions of the Institution of the Rubber Industry*, 22, 82–110.
- Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey P. A. (1983), *Graphical Methods for Data Analysis*, New York: Chapman and Hall.
- Cleveland, W. S. (1985), *The Elements of Graphing Data*, Summit: Hobart Press.
- (1993), *Visualizing Data*, Summit: Hobart Press.
- Cook, D., Hofmann, H., Lee, E., Yang, H., Nikolau, B., and Wurtele, E. (in press), "Visual Methods for Data from Two Factor Single-Replicate Gene Expression Studies," *Journal of Computational and Graphical Statistics*, submitted for publication. Available on-line at www.public.iastate.edu/~dicook/papers/Microarray/paper.pdf.
- Cook, D., Swayne, D.F., Buja, A., and Hofmann, H., (2004), "Interactive and Dynamic Graphics for Data Analysis: With Examples Using R and GGobi," in preparation. Available on-line at www.public.iastate.edu/~dicook/ggobi-book/ggobi.html.
- Craig, B. A., Black, M. A., and Doerge, R. W. (2003), "Gene Expression Data: the Technology and Statistical Analysis," *Journal of Agricultural, Biological, and Environmental Statistics*, 8, 1–28.
- Daniel, C., and Wood, F. (1980), *Fitting Equations to Data* (2nd ed.), New York: Wiley.
- Davies, O. L. (1947), *Statistical Methods in Research and Production*, London: Oliver and Boyd.
- Davies, O. L., Box, G. E. P., Cousins, W. R., Himsworth, F. R., Kenney, H., Milbourn, M., Spendley, W., and Stevens, W. L. (1957), *Statistical Methods in Research and Production* (3rd ed.), London: Oliver and Boyd.
- Donoho, D. L., Donoho, D. L., and Gasko, M. (1985), *MACSPIN Graphical Data Analysis Software*, Austin: D² Software.
- Fishkeller, M. A., Friedman, J. H., and Tukey, J. W. (1974), "PRIM-9: An Interactive Multidimensional Data Display and Analysis System," in *Proceedings of the Pacific ACM Regional Conference*.
- Fowlkes, E. B. (1971), "User's Manual for an On-line Interactive System for Probability Plotting on the DDP-224 Computer," Technical Memorandum, Bell Laboratories, Murray Hill, NJ.
- Hand, D. J., Daly, F., Lunn, A. D., McConway, K., and Ostrowski, E. (eds.) (1994), *A Handbook of Small Data Sets*, London: Chapman and Hall.

- Nguyen, D., Bulak, A., Naisyin, W., and Carroll, R. (2002), "DNA Microarray Experiments: Biological and Technological Aspects," *Biometrics*, 58, 707–717.
- Preece, D. A. (1987), "Good Statistical Practice," *The Statistician*, 36, 397–408.
- Reaven, G. M., and Miller, R. G. (1979), "An Attempt to Define the Nature of Chemical Diabetes Using Multidimensional Analysis," *Diabetologia*, 16, 17–24.
- Scott, D. W. (2003), "The Case for Statistical Graphics," *AMSTAT News*, September 2003, 20–22.
- Tukey, J. W. (1973), "Some Thoughts on Alternative Graphic Displays," Technical Report 45, Series 2, Department of Statistics, Princeton University.
- Tufte, E. R. (1990), *Envisioning Information*, Cheshire: Graphics Press.
- Unwin, A. (1992), "How Interactive Graphics will Revolutionize Statistical Practice," *The Statistician*, 4, 365–369.
- Unwin, A. (1999), "Requirements for Interactive Graphics Software for Exploratory Data Analysis," *Computational Statistics*, 14, 7–22.
- Velleman, P. F. (1999), *Data Desk* (Version 6.1), Ithaca: Data Description.