

Quick and Easy Visualization of Longitudinal data with the WEBFRAME graphics device

Kevin P. Delaney MPH, Centers for Disease Control and Prevention, Atlanta, GA

Abstract:

Data Visualization is an important step in getting to know your data. This is particularly true when you have longitudinal data, repeated measurements on the same subjects over time. A useful tool in visualization of longitudinal data is the "Trellis graph" a graph of the data separated into multiple graphs by a conditional variable. In the case of longitudinal data, this conditional variable is often the individual participant. Several SAS competitors have made the ability to visualize data using "Trellis graphs" an important selling point for their product. Trying to do something similar with traditional SAS/GRAPH methods is not easily done. However, Version 8 of SAS brought us the little known WEBFRAME graphics device which can be used to create a modified version of the Trellis graph with ease. We take a quick look at this device as a simple way to get look at your data, with particular attention to longitudinal data. Although there is a complicated PROC GREPLAY included as an example, the purpose of this presentation is to show you how refreshingly easy the WEBFRAME graphics device is to use, even for a SAS/GRAPH novice.

Introduction

Any Statistician will tell you that the first step in data analysis is to get to know your data. This typically means looking at basic frequency distributions of important discrete variables as well as measures of both central tendency (mean, median) and spread (Standard deviation, percentiles) of continuous variables. In studies that involve longitudinal data collection, defined here as measurements collected on the same subject (and therefore usually correlated) over time, it is important to include descriptions of the periodicity and spread of these measurements in your descriptions of the data. Thus, characteristics of the data such as the number of time points included for each subject, the total length of follow-up time and some description of the spread between time points if they are unequally spaced are often included in the beginning phases of analysis. These descriptive statistics are essential for guiding your analysis plan.

Often it is useful to look at not only tables of these important characteristics of your data, but also to create graphs and charts of your data in order to actually SEE what it looks like. Normality plots and histograms should be included in the descriptions of important variables to both describe the data and help guide the analysis. For longitudinal data it is important to include time as an additional variable in your graphic representations. This is typically accomplished by graphing variables of interest over time, with Time as the X-axis and the variable of interest on the Y-axis. Such graphs are often referred to as 'profile plots' and may include data for individuals over time and/or trend lines for the data as a whole.

There is one major drawback to a traditional "Profile plot" of your data. If you plot data for each subject as a separate line on the graph, and you have more than 50 or so subjects you will end up with a graph that looks like a plate of spaghetti in two dimensions. The SAS Institute "Analysis of Longitudinal Data" course includes a dataset of CD4 counts over time collected from 369 AIDS patients with 1 to 12 follow-up time points. In that course one of the first exercises is to graph a profile plot of the CD4 counts over time, with a line for each patient and an overall trend line for the data, that graph is shown below as Figure 1 (The Code for generating this graph, is included in the Appendix as Example 1).

While this graph provides an overall trend to the data, the sheer volume of data presented on it leaves several questions unanswered.

- 1) How many participants are included?
- 2) How many have at least 3 follow-up timepoints?
- 3) How long is the average follow-up for these participants?
- 4) Are there outliers hidden in this puddle of information on a page?
- 5) Are there individual patients that, while not diagnostically considered to be outliers in the data, might still exhibit an interesting trend that is hidden in the aforementioned mush?

“Trellis Graphs and PROC GREPLAY”

With data like these, it is possible to visualize the data in way that can provide answers to these questions. You can plot a separate graph of CD4 counts over time for each individual client in the data set. This is a very simple form of the graphic technique known as a “Trellis Graph” where you plot a separate graph for a “conditional variable” that may be affecting the relationship between your X and Y variables. In this case, in order to see the effect of an individual on our longitudinal data, we are plotting separate graphs for each person.

You COULD use PROC GREPLAY to put together pages of graphs of individual clients. For instance something like the code in Example 2 (See Appendix) could be used to fit as many as 20 graphs of individual profiles on a single page in the PDF destination. This code produces the somewhat useful set of graphs shown in Figure 2. But, remember in this instance there are 369 individuals included in these data, which means 19 pages of these graphs, and it would require a rather complicated Macro to produce these 20 plots per page for the different number of individuals likely to be found in each of your longitudinal datasets.

The Point: Visualizing your DATA with the WEBFRAME Graphics DEVICE DRIVER

However, the functionality to produce a variation of a traditional “Trellis graph” is available in SAS, and is particularly easy to use for the purpose of visualizing longitudinal data for each client or participant in your study. The WEBFRAME graphics device driver makes visualizing data broken up by person or any other conditional variable easy. When you use the WEBFRAME device SAS generates three different types of files (A screen shot of the GSFNAME folder with all these files is shown in Figure 3):

- 1) A GIF image of each graph you generate while the device is open,
- 2) A “thumbnail” version of each of these graphs
- 3) A Series of HTML pages linked together in a file called Index.html.

The Index.html file has two frames (See Figure 4), one with all of the graphs generated in ‘thumbnail’ form for easy scrolling through the individual graphs, and one with a larger version of (initially) the first graph generated. If you click on a thumbnail (or flip between Figures 4 and 5), the graph in the “viewing window” switches to the full size version of the graph you clicked on. As you can see, the WEBFRAME device makes it very easy to look at the differences in the data that become apparent when you are able to produce a series of graphs BY a “conditional variable”, and it can be done in three easy steps!

*Example 3: Using the WEBFRAME device driver;

*1) Specify a folder where SAS should write out your graphs;

filename outfiles "C:\Graph Home";

*2) Specify the WEBFRAME device driver on the DEVICE option,
and use the filename identified in 1) as the destination
for the graphs with the GSFNAME option;

options device=webframe gsfname=outfiles;

proc gplot data=sasuser.aids;

*3) Use a BY statement with you 'Conditional Variable' (you might have
to SORT the dataset.) In this case we are using ID to get a
separate graph for each ID.;

by id;

plot cd4*time/Name="Client";

run;

quit;

options reset=all;

The ability to see each study participant's data in their own graph, and to simultaneously compare it to each of other participant's graphs is a very powerful data visualization tool. In our example of the AIDS cohort data, it allows us to identify persons who have only 2 or 3 data points, persons who have higher than average CD4 counts throughout follow-up, and even "rapid progressers" whose CD4 counts decline very rapidly after seroconversion.

Unlike the PROC GREPLAY solution, in this case the WEBFRAME graphics device does all the work for you. Its simplicity is the key to its utility. Conceivably, you could preprocess your data with little effort, in order to create a BY variable that corresponded to any conditional variable or combination of conditional variables that you choose. You could then use BY group processing to create any number of these modified Trellis Graph's using the WEBFRAME device.

Conclusion

I am not sure the developers of the SAS WEBFRAME device know what a powerful tool they have created. The ease with which you can visualize your data broken out by any number of conditional variables really is surprising, especially when you consider you are using SAS/GRAPH to do it and not an unmentioned competitors special "Trellis Graph" utilities. Since it is so easy to use, I encourage you to play with it, and get a good look at your data, particularly your longitudinal data. Just like you can never know your data too well, you can never know your SAS/GRAPH tools too well. Hopefully this paper will inspire you to start using this little known tool more often!

References

SAS Institute, Inc. (1999) SAS/GRAPH Reference, Version 8. Cary, NC. SAS Institute, Inc. Particularly the Chapter on PROC TEMPLATE

SAS Institute, Inc. (2002) Analysis of Longitudinal Data Course Notes. SAS Institute, Inc. Cary, NC

Contact Information:

Kevin P. Delaney MPH
Centers for Disease Control and Prevention
Atlanta, GA
KDelaney@cdc.gov

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies

Figure 1: Profile plot, with TOO MUCH DATA

Individual Profiles with the Average Trend Line

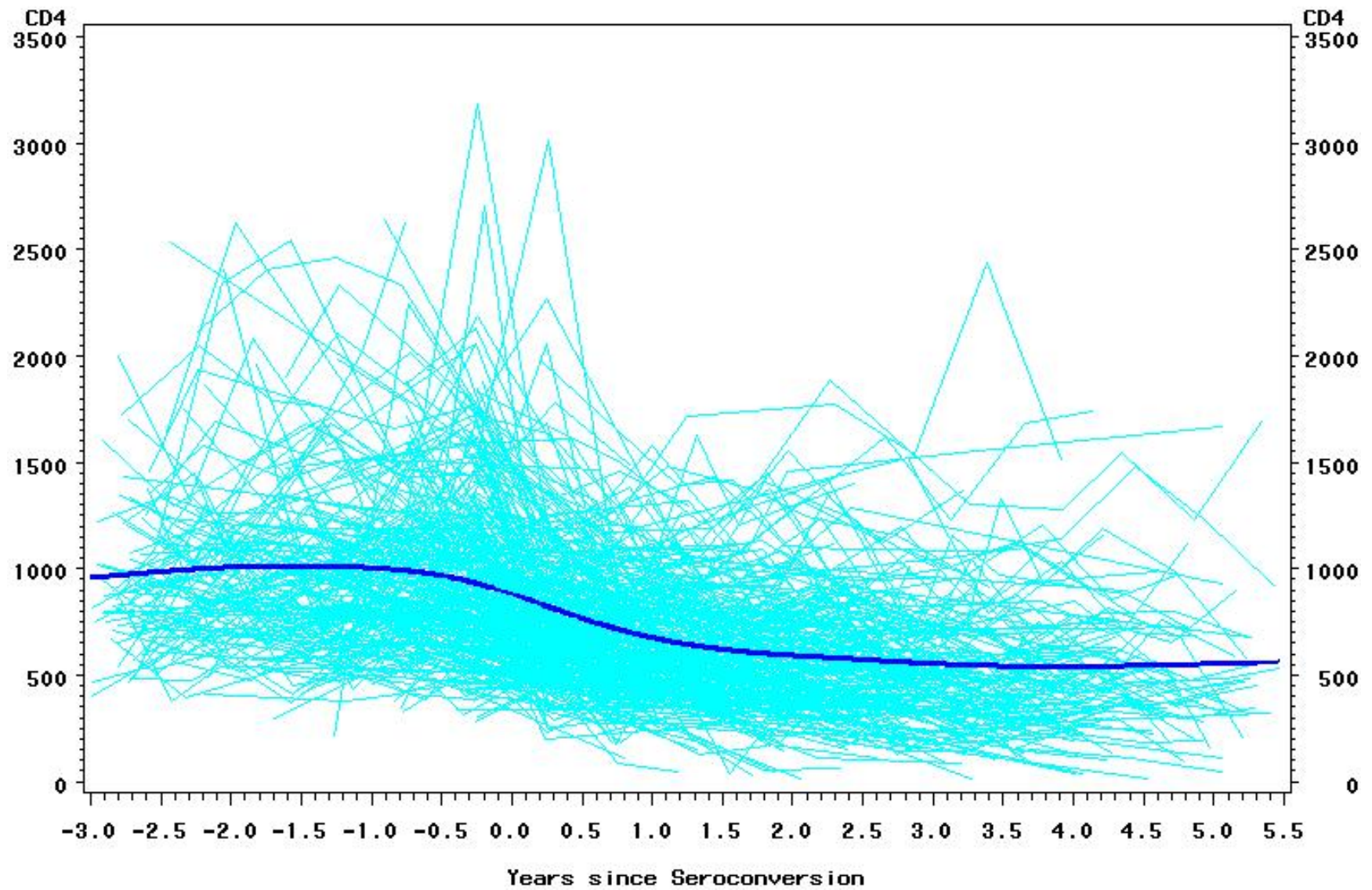


Figure 2: "Trellis plot" of individual data points from PROC GREPLAY

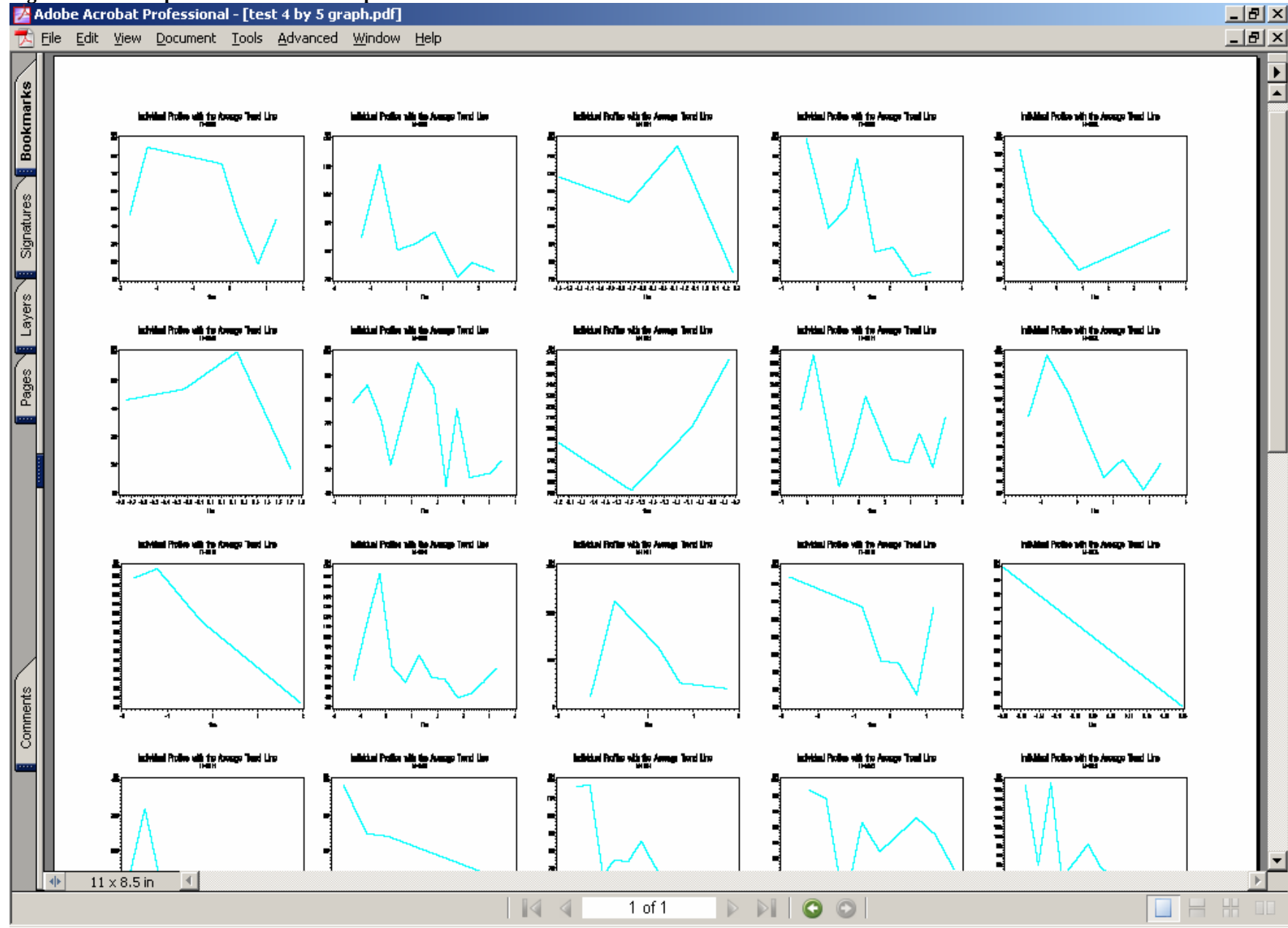


Figure 3: A look at the files created by the WEBFRAME graphics device (in the GSFNAME folder)

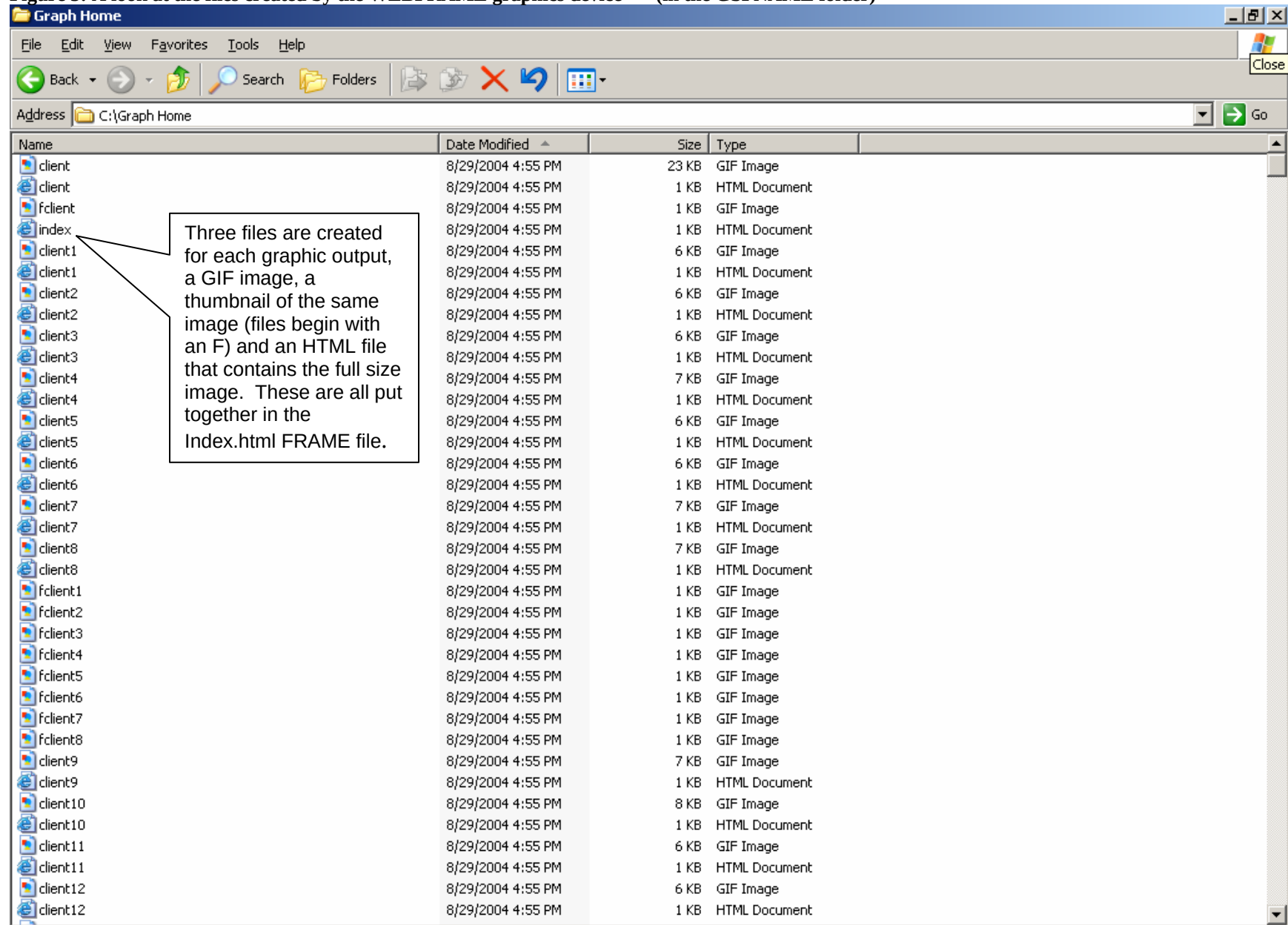


Figure 4: Open view of Index.html

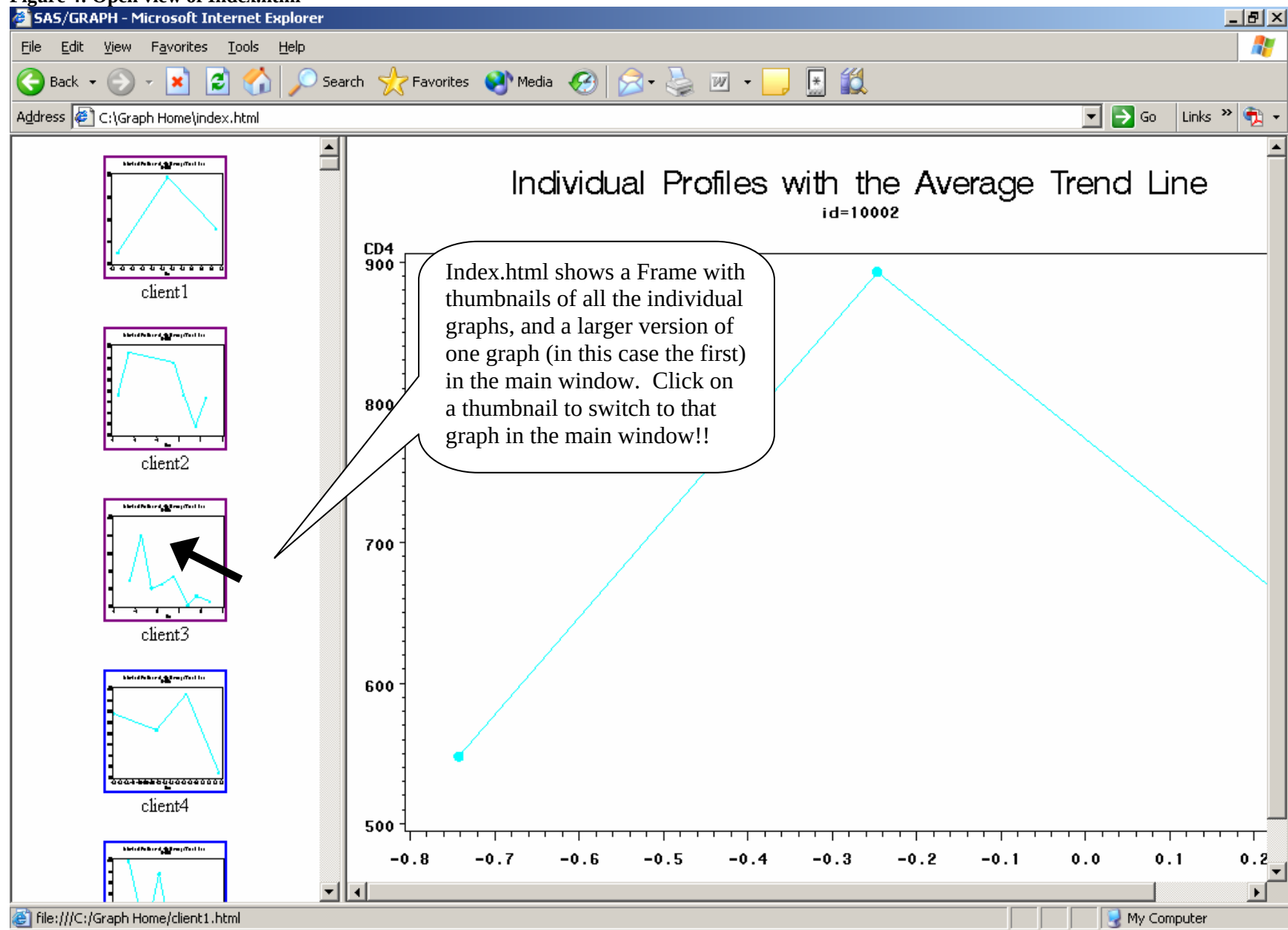
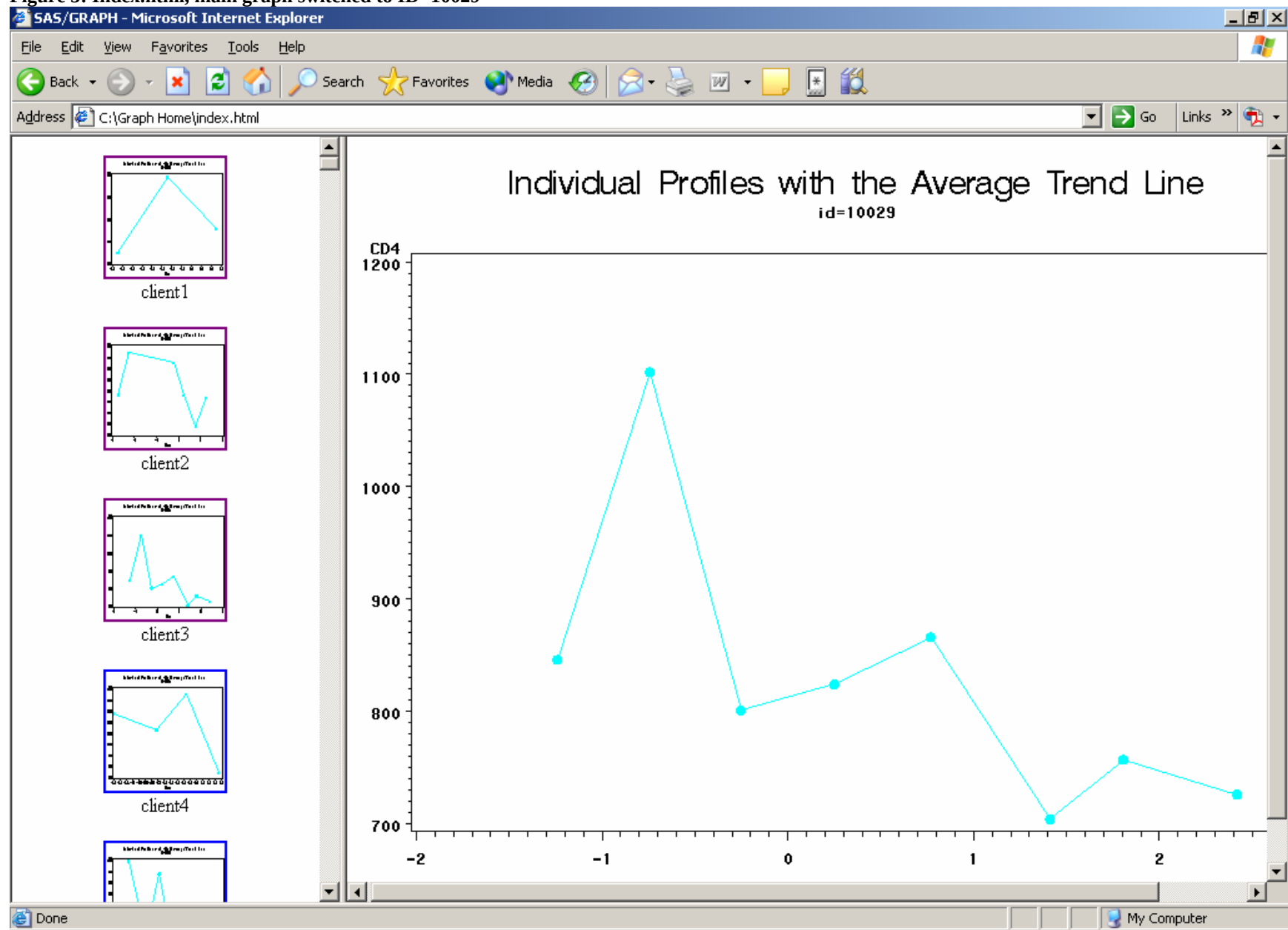


Figure 5: Index.html, main graph switched to ID=10029



Appendix: SAS code for examples 1 & 2

*All these examples use the AIDS dataset from the SAS Longitudinal Data Analysis Class;
 *To get the data, you can either take the class or contact me and I will see what I can do.
 Personally, if I were you, I would adapt these examples for use with your own Longitudinal data...;
 *Example 1: Profile plot with TOO MUCH data!;

```
goptions reset=all;
proc gplot data=sasuser.aids;
  plot cd4*time=id
    / haxis = -3 to 5.5 by 0.5
      vaxis = 0 to 3500 by 500;
  plot2 cd4*time
    / haxis = -3 to 5.5 by 0.5
      vaxis = 0 to 3500 by 500;
  symbol1 v=none repeat=369 i=join color=cyan;
  symbol2 v=none i=sm50s color=blue width=3;
  label time='Years since Seroconversion';
  title 'Individual Profiles with the Average Trend Line';
run;
quit;
```

*Example 2: Create a "Trellis Plot" with Person as the conditional variable,
 using PROC GREPLAY to put 20 graphs on a page;
 *Problems with this method include:
 It's too hard to read these little graphs and you can't make them any bigger
 Generating these graphs for a different number of persons or other conditional variables
 every time would require a flexible and complicated Macro
 When your boss says, 20 is too many I want 15 per page, redoing the GREPLAY is not very easy, either...
 ;

```
*Use PROC CATALOG to get rid of old graphs;
proc catalog cat=work.gseg kill;
quit;
*Generate a separate graph for each person with a BY statement (make sure the data is
  already sorted!);
proc gplot data=sasuser.aids;
  by id;
  plot cd4*time/Name="Client";
run;
quit;
```

```
goptions rotate=landscape hsize=11 vsize=8.5 device=pdfc;
```

```
options orientation=landscape;
ods pdf file="C:\test 4 by 5 graph.pdf" ;
*Spend all day figuring out dimensions of PROC GREPLAY to fit 4 rows of 5 graphs
  on a landscape page in a PDF;
proc greplay tc=work.tempcat nofs;
```

tdef fiveby4 des='Cram 20 graphs in one page'

1/ulx=2.5 uly=97.25
 urx=19.5 ury=97.25
 llx=2.5 lly=75.25
 lrx=19.5 lry=75.25

2/ulx=21 uly=97.25
 urx=38 ury=97.25
 llx=21 lly=75.25
 lrx=38 lry=75.25

3/ulx=40.5 uly=97.25
 urx=57.5 ury=97.25
 llx=40.5 lly=75.25
 lrx=57.5 lry=75.25

4/ulx=60 uly=97.25
 urx=77 ury=97.25
 llx=60 lly=75.25
 lrx=77 lry=75.25

5/ulx=79.5 uly=97.25
 urx=96.5 ury=97.25
 llx=79.5 lly=75.25
 lrx=96.5 lry=75.25

6/ulx=2.5 uly=72.75
 urx=19.5 ury=72.75
 llx=2.5 lly=50.75
 lrx=19.5 lry=50.75

7/ulx=21 uly=72.75
 urx=38 ury=72.75
 llx=21 lly=50.75
 lrx=38 lry=50.75

8/ulx=40.5 uly=72.75
 urx=57.5 ury=72.75
 llx=40.5 lly=50.75
 lrx=57.5 lry=50.75

9/ulx=60 uly=72.75
 urx=77 ury=72.75
 llx=60 lly=50.75
 lrx=77 lry=50.75

10/ulx=79.5 uly=72.75
 urx=96.5 ury=72.75
 llx=79.5 lly=50.75
 lrx=96.5 lry=50.75

11/ulx=2.5 uly=48.25
 urx=19.5 ury=48.25
 llx=2.5 lly=26.25
 lrx=19.5 lry=26.25

12/ulx=21 uly=48.25
 urx=38 ury=48.25
 llx=21 lly=26.25
 lrx=38 lry=26.25

13/ulx=40.5 uly=48.25
 urx=57.5 ury=48.25
 llx=40.5 lly=26.25
 lrx=57.5 lry=26.25

14/ulx=60 uly=48.25
 urx=77 ury=48.25
 llx=60 lly=26.25
 lrx=77 lry=26.25

15/ulx=79.5 uly=48.25
 urx=96.5 ury=48.25
 llx=79.5 lly=26.25
 lrx=96.5 lry=26.25

16/ulx=2.5 uly=23.75
 urx=19.5 ury=23.75
 llx=2.5 lly=1.75
 lrx=19.5 lry=1.75

17/ulx=21 uly=23.75
 urx=38 ury=23.75
 llx=21 lly=1.75
 lrx=38 lry=1.75

18/ulx=40.5 uly=23.75
 urx=57.5 ury=23.75
 llx=40.5 lly=1.75
 lrx=57.5 lry=1.75

19/ulx=60 uly=23.75
 urx=77 ury=23.75
 llx=60 lly=1.75
 lrx=77 lry=1.75

20/ulx=79.5 uly=23.75
 urx=96.5 ury=23.75
 llx=79.5 lly=1.75
 lrx=96.5 lry=1.75

```
;
```

```
template fiveby4;
```

```
igout work.gseg;
```

```
gout work.gseg;
```

```
treplay
```

```
1:client1  
2:client2  
3:client3  
4:client4  
5:client5  
6:client6  
7:client7  
8:client8  
9:client9  
10:client10  
11:client11  
12:client12  
13:client13  
14:client14  
15:client15  
16:client16  
17:client17  
18:client18  
19:client19  
20:client20  
name="5_BY_4";
```

```
quit;
```

```
ods pdf close;
```