Multivariate Analysis Using Parallel Coordinates

Stephen Few

September 12, 2006

This article discusses parallel coordinates, an approach to analyzing multivariate data using data visualization techniques.

This article is part of a series that I began in July of this year with the article entitled "An Introduction to Visual Multivariate Analysis." In that initial article, I provided an overview of several approaches to analyzing multivariate data using visualization techniques. In this article, I am featuring an approach called *parallel coordinates*.

The first time that I saw a parallel coordinates visualization, I almost laughed out loud. My initial impression was "How absurd!" I couldn't imagine how anyone could make sense of the dense clutter caused by hundreds of overlapping lines (see Figure 1). This certainly isn't a chart that you would present to the board of directors or place on your Web site for the general public. In fact, the strength of parallel coordinates isn't in their ability to communicate some truth in the data to others, but rather in their ability to bring meaningful multivariate patterns and comparisons to light when used interactively for analysis.

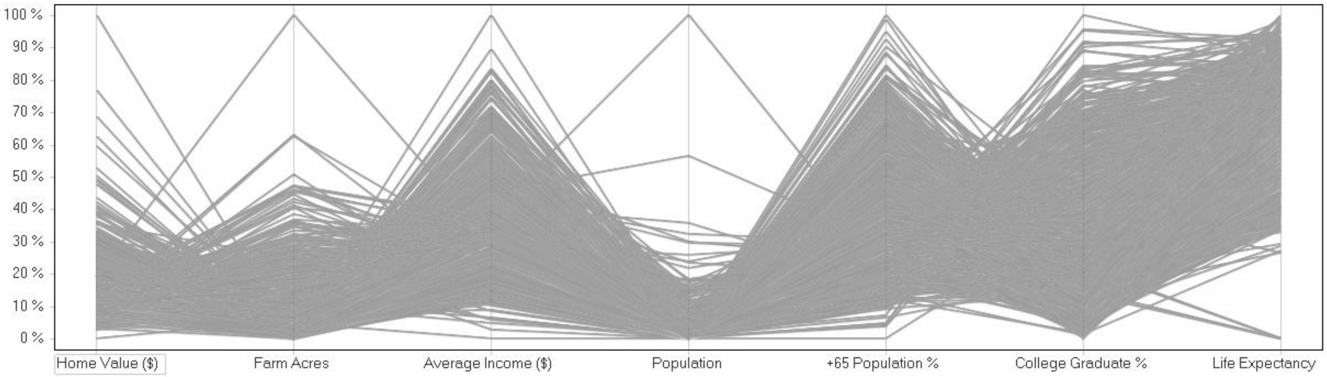


Figure 1: A parallel coordinates display that measures several aspects of U.S. counties.

Reading Parallel Coordinates

To recognize the worth of a parallel coordinates display, you cannot think of it as a normal line graph. Lines are predominantly used to encode time-series data. The up and down slopes of the lines indicates change through time from one value to the next. The lines in parallel coordinate displays, however, don't indicate change. Instead, a single line in a parallel coordinates graph connects a series of values - each associated with a different variable - that measure multiple aspects of something, such as a person, product, or country. The example in Figure 1 consists of 3,138 lines: one for each county in the United States. This graph has seven vertical axes arranged from left to right along the X-axis, each for a different variable that measures some aspect of U.S. counties, including median home value, the number of farm acres, the average per capita income, and so on. Notice that the units of measure differ among these variables, including dollars, counts, acres, percentages, and years. In this particular example, the scales for these independent variables have been normalized as percentages, with the highest value at the top (100%) and the lowest at the bottom (0%). For example, when this data was collected a few years ago, median home values ranged from $20,100 in McPherson County, South Dakota at the low end to $750,000 in Pitkin County, Colorado at the high end.

Figure 2 displays the same set of data, but this time the line representing a single county has been highlighted. I selected Alameda County in California, where I live, to see its multivariate profile.

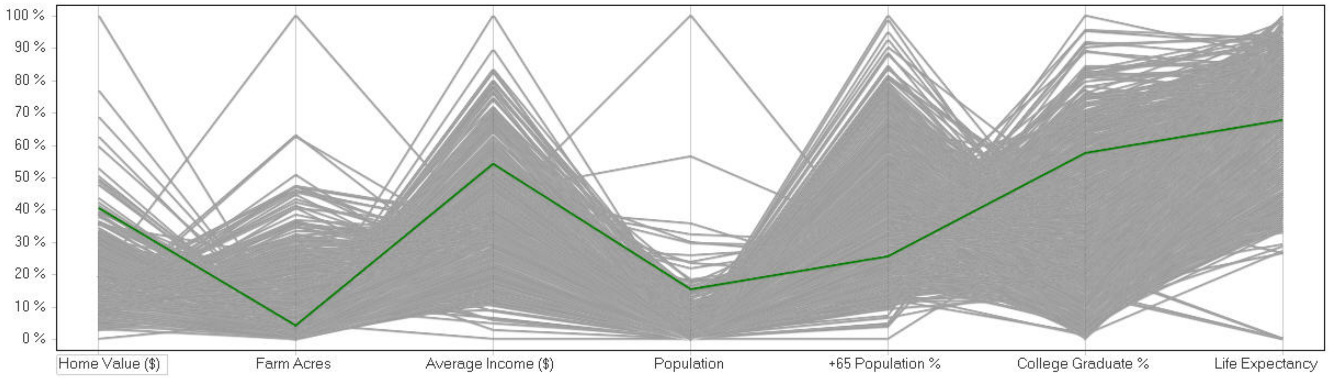


Figure 2: Alameda County in the state of California has been highlighted.

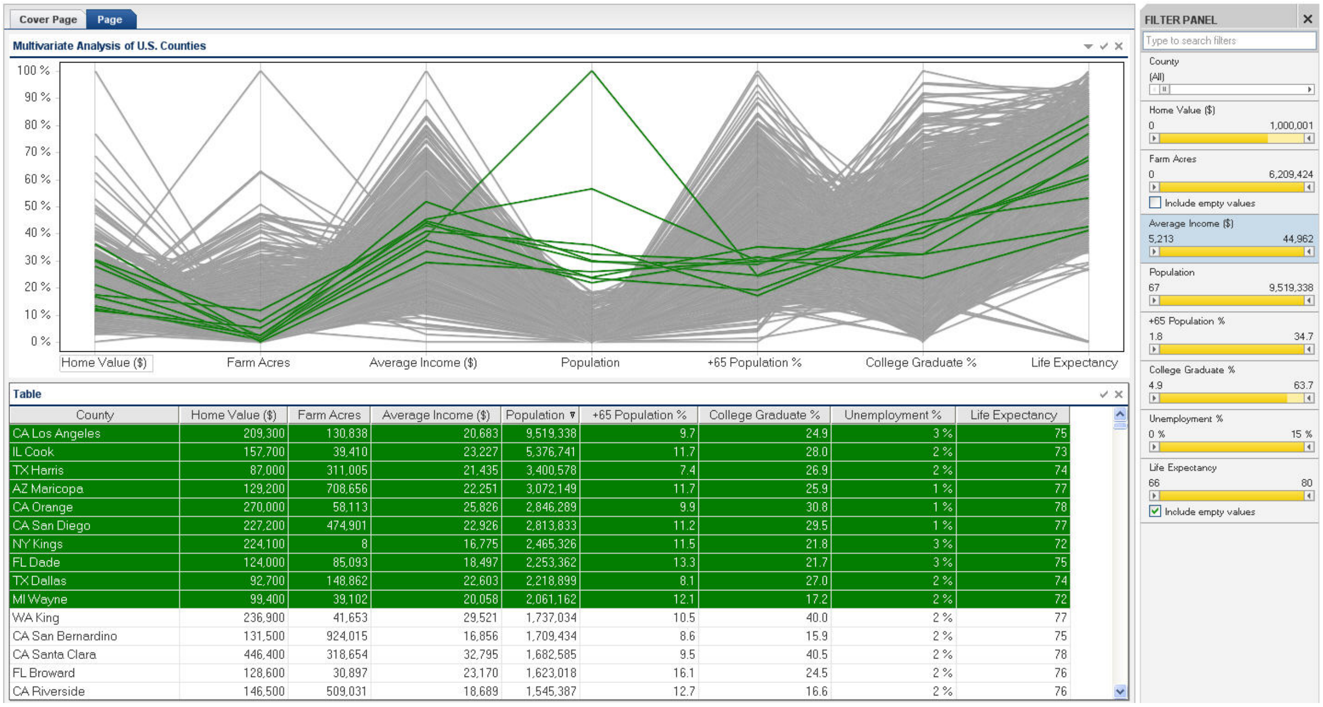
In examining this Alameda County profile, we must be careful to read nothing of significance into the slope of each line segment or the overall pattern formed by the line as a whole. The slopes and overall pattern would look completely different if I rearranged the order of the variables. Instead, we should read the variables one by one to construct a composite profile of Alameda County. In doing so, because we can see Alameda County in the context of all counties, we can quickly determine that home values are higher than average but only about 40% of that of the county with the highest value, the number of farm acres is much lower than average, income level is higher than most but only about 55% of the county with the highest value, and so on.

The Big Picture

Look again at Figure 1 to see if you can discern anything meaningful from its clutter of overlapping lines. Don't approach it as you would a normal line graph, but rather as a multivariate overview of 3,138 counties. When doing multivariate analysis, the big picture is usually where you want to start, for meaningful observations, believe it or not, can be gleaned from the clutter. A display with this much data cannot be used to explore the details, but it can be used to search for predominant patterns and exceptions. For example, we can tell that all but a few counties have populations that are 20% or less than the county with the highest population. The county with the highest population - Los Angeles - stands out as a clear exception with approximately twice the population of the next highest county - Cook County, Illinois. Several other exceptions also assert themselves, such as the fact that a few counties have life expectancies that are much lower than most (all are in South Dakota). By starting the analytical process with the big picture, we can then dig down into the predominant patterns and exceptions that catch our eyes.

Useful Ways to Complement and Interact with Parallel Coordinates

The examples that we've seen so far were created using Spotfire DXP, which enables us to complement the parallel coordinates graph with other displays and several useful means to interact with the data. In Figure 3, I've expanded the screenshot to show controls that can be used to filter the data (to the right of the parallel coordinate graph) and a table that provides precise details about the information that appears in the graph. In this case, I decided to look at the profiles for the 10 counties with the largest populations, which I accomplished by sorting the table by population from highest to lowest and selecting the top ten rows. By highlighting these rows in the table, the corresponding lines in the graph were automatically highlighted as well, which makes it easy to see that counties with the highest populations all consist of relatively low farm acreage.



County	Home Value ($)	Farm Acres	Average Income ($)	Population	+65 Population %	College Graduate %	Unemployment %	Life Expectancy
CA Los Angeles	209,300	130,838	20,683	9,519,338	9.7	24.9	3 %	75
IL Cook	157,700	39,410	23,227	5,376,741	11.7	28.0	2 %	73
TX Harris	87,000	311,005	21,435	3,400,578	7.4	26.9	2 %	74
AZ Maricopa	129,200	708,656	22,251	3,072,149	11.7	25.9	1 %	77
CA Orange	270,000	58,113	25,826	2,846,289	9.9	30.8	1 %	78
CA San Diego	227,200	474,901	22,926	2,813,833	11.2	29.5	1 %	77
NY Kings	224,100	8	16,775	2,465,326	11.5	21.8	3 %	72
FL Dade	124,000	85,093	18,497	2,253,362	13.3	21.7	3 %	75
TX Dallas	92,700	148,862	22,603	2,218,899	8.1	27.0	2 %	74
MI Wayne	99,400	39,102	20,058	2,061,162	12.1	17.2	2 %	72
WA King	236,900	41,653	29,521	1,737,034	10.5	40.0	2 %	77
CA San Bernardino	131,500	924,015	16,856	1,709,434	8.6	15.9	2 %	75
CA Santa Clara	446,400	318,654	32,795	1,682,585	9.5	40.5	2 %	78
FL Broward	128,600	30,897	23,170	1,623,018	16.1	24.5	2 %	76
CA Riverside	146,500	509,031	18,689	1,545,387	12.7	16.6	2 %	76

County
(All)

Home Value ($)
0 1,000,001

Farm Acres
0 6,209,424
☐ Include empty values

Average Income ($)
5,213 44,962

Population
67 9,519,338

+65 Population %
1.8 34.7

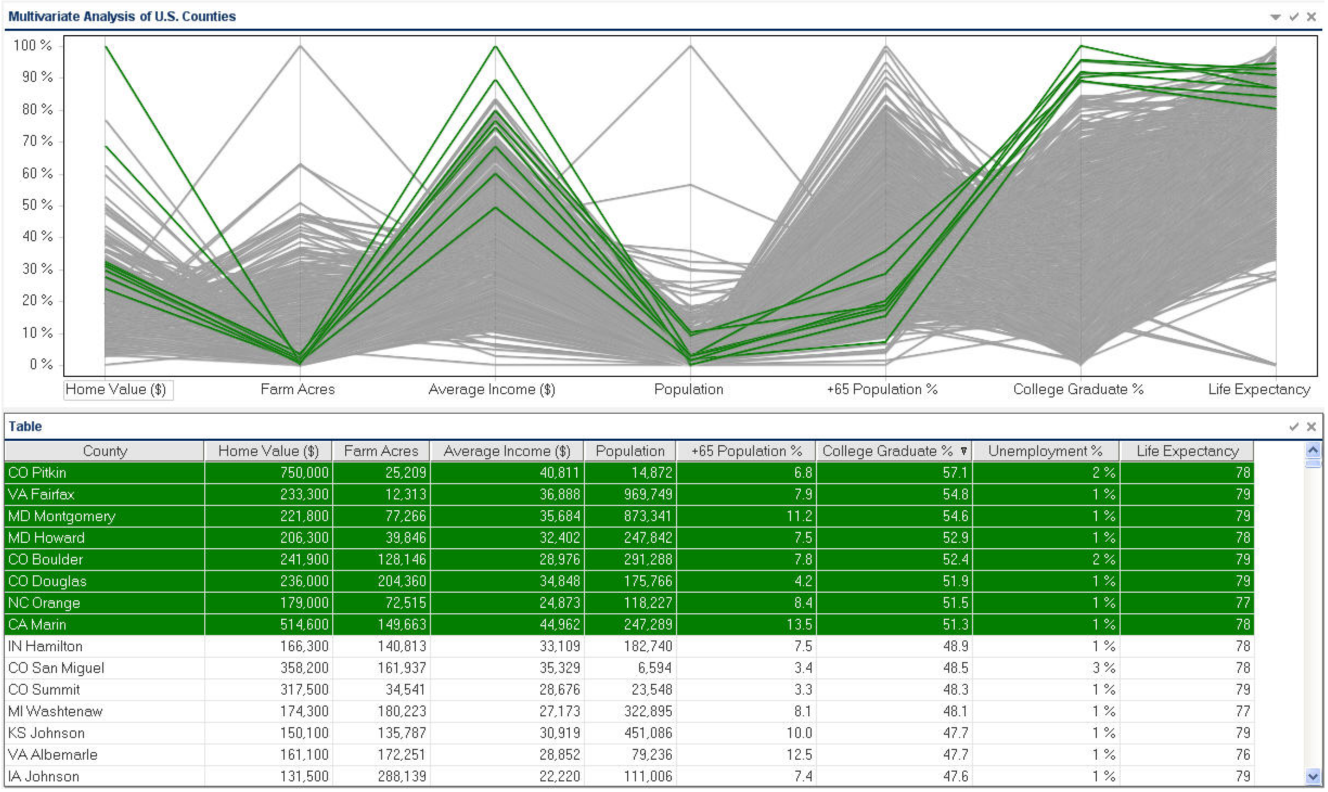
College Graduate %
4.9 63.7

Unemployment %
0 % 15 %

Life Expectancy
66 80
☑ Include empty values

Figure 3: The 10 counties with the largest populations.

Another way that I can easily highlight items is to simply draw a rectangle around values in the graph itself that interest me. In Figure 4, you can see the results of drawing a rectangle around the highest values on the *College Graduate %* axis. Given the resulting view, it only takes a moment to notice that counties with the highest percentages of college graduates all have very few acres of farmland, higher than average incomes, relatively small populations, and high life expectancies.



County	Home Value ($)	Farm Acres	Average Income ($)	Population	+65 Population %	College Graduate %	Unemployment %	Life Expectancy
CO Pitkin	750,000	25,209	40,811	14,872	6.8	57.1	2 %	78
VA Fairfax	233,300	12,313	36,888	969,749	7.9	54.8	1 %	79
MD Montgomery	221,800	77,266	35,684	873,341	11.2	54.6	1 %	79
MD Howard	206,300	39,846	32,402	247,842	7.5	52.9	1 %	78
CO Boulder	241,900	128,146	28,976	291,288	7.8	52.4	2 %	79
CO Douglas	236,000	204,360	34,848	175,766	4.2	51.9	1 %	79
NC Orange	179,000	72,515	24,873	118,227	8.4	51.5	1 %	77
CA Marin	514,600	149,663	44,962	247,289	13.5	51.3	1 %	78
IN Hamilton	166,300	140,813	33,109	182,740	7.5	48.9	1 %	78
CO San Miguel	358,200	161,937	35,329	6,594	3.4	48.5	3 %	78
CO Summit	317,500	34,541	28,676	23,548	3.3	48.3	1 %	79
MI Washtenaw	174,300	180,223	27,173	322,895	8.1	48.1	1 %	77
KS Johnson	150,100	135,787	30,919	451,086	10.0	47.7	1 %	79
VA Albemarle	161,100	172,251	28,852	79,236	12.5	47.7	1 %	76
IA Johnson	131,500	288,139	22,220	111,006	7.4	47.6	1 %	79

Figure 4: The counties with the highest percentages of college graduates have been highlighted.

It is often helpful to separate clusters of similar data into separate graphs to more easily focus on specific groups independent of the others and to compare their multivariate profiles. In Figure 5, to pursue an interest in the relationship between the percentage of college graduates and the other variables, I used convenient functionality in Spotfire DXP to divide the data into five groups (or bins) based on the percentage of college graduates and to place each group into a separate graph. The top graph displays counties with the lowest percentage of college graduates and in the bottom graph we see those with the highest percentages. A quick comparison of these graphs reveals that counties with the lowest percentages of college graduates also have the lowest home values as well as widely ranging percentages of elderly residents compared to counties with the highest percentages of college graduates. Another difference between these five groups that surfaces when viewed in this fashion is that the distribution of values for each variable except home value and population tends to narrow with each graph, beginning with the top graph (lowest percentage of college graduates), which displays a broad distribution of values across most variables, and proceeding down to the bottom graph (highest percentage of college graduates), which displays a relatively narrow distribution of values for each variable. In other words, greater percentages of college graduates appear to correspond to greater homogeneity among the people in that county.

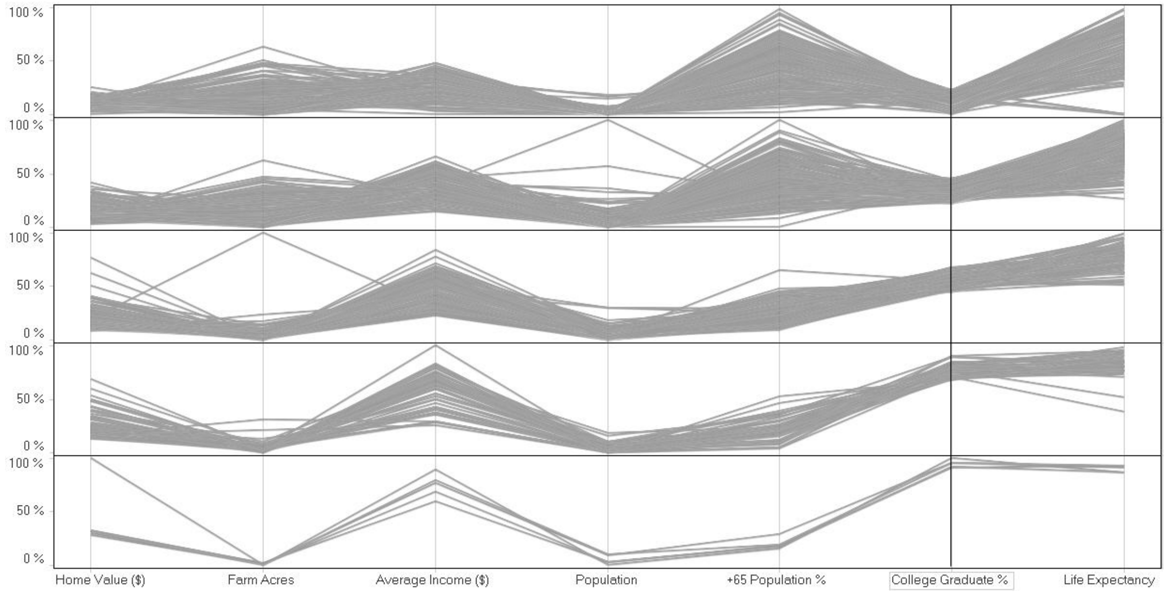
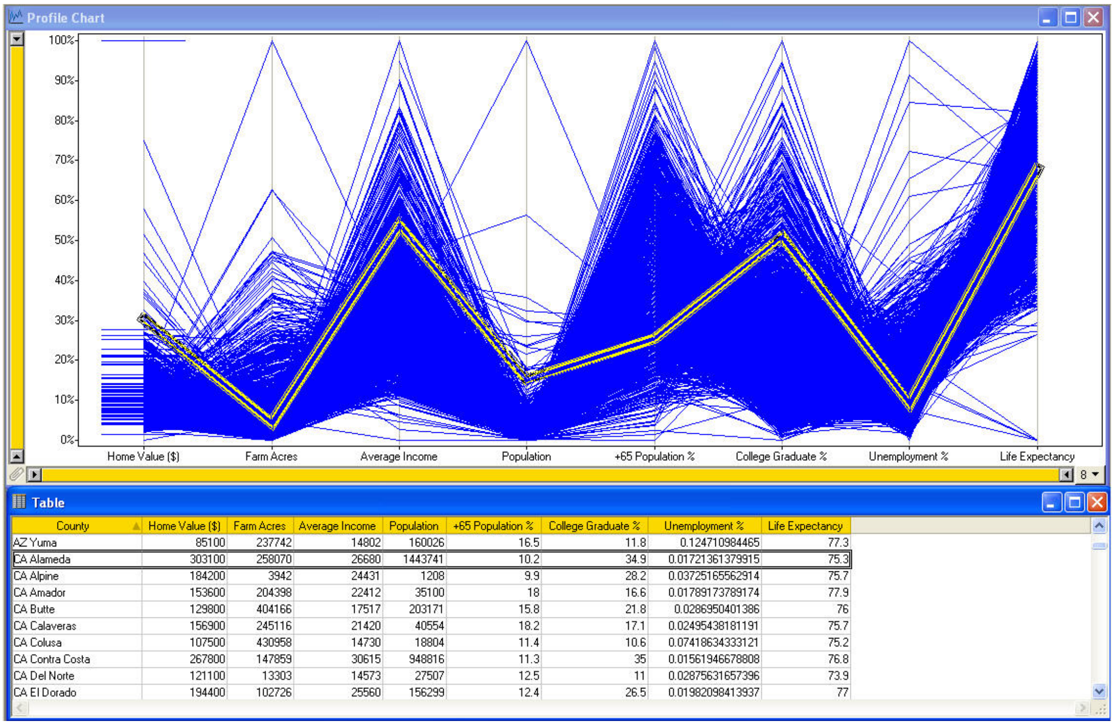


Figure 5: This display consists of five parallel coordinates graphs based on the percentage of college graduates.

Searching for Similar Profiles

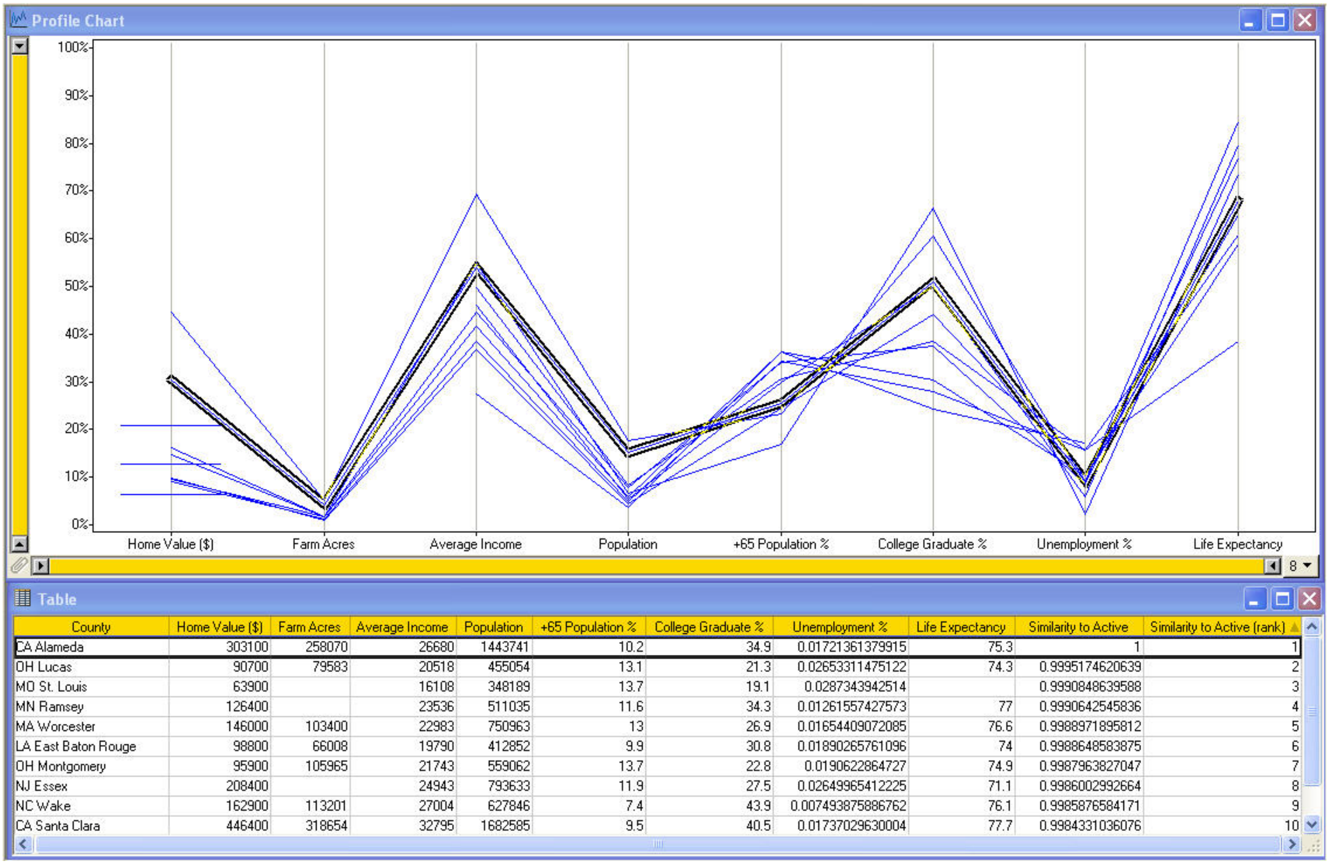
Another useful task when exploring multivariate data involves searching for entities with a particular multivariate profile—either one that is exhibited by a particular entity (such as a county in the examples above) or one that you imagine might be interesting. To illustrate how this works, I've switched over to Spotfire Decisionsite to access this functionality, but am still examining the same set of data. Notice in Figure 6 that I've selected Alameda County once again (the highlighted line), which I'll use as the model profile for my pattern search.



County	Home Value ($)	Farm Acres	Average Income	Population	+65 Population %	College Graduate %	Unemployment %	Life Expectancy
AZ Yuma	85100	237742	14802	160026	16.5	11.8	0.124710984465	77.3
CA Alameda	303100	258070	26680	1443741	10.2	34.9	0.01721361379915	75.3
CA Alpine	184200	3942	24431	1208	9.9	28.2	0.03725165562914	75.7
CA Amador	153600	204398	22412	35100	18	16.6	0.01789173789174	77.9
CA Butte	129800	404166	17517	203171	15.8	21.8	0.0286950401386	76
CA Calaveras	156900	245116	21420	40554	18.2	17.1	0.02495438181191	75.7
CA Colusa	107500	430958	14730	18804	11.4	10.6	0.07418634333121	75.2
CA Contra Costa	267800	147859	30615	948816	11.3	35	0.01561946678808	76.8
CA Del Norte	121100	13303	14573	27507	12.5	11	0.02875631657396	73.9
CA El Dorado	194400	102726	25560	156299	12.4	26.5	0.01982098413937	77

Figure 6: Preparing to search for counties with profiles that are similar to Alameda County.

After running the search for counties with similar profiles and viewing the results, I selected the 10 counties most similar to Alameda County and removed all but them from the display to eliminate distractions. You can see the results in Figure 7, which shows the 10 counties in the parallel coordinates graph along with Alameda County, which is highlighted. These counties also appear in the table, which now includes two new columns that were produced by the search operation: "Similarity to Active," which measures their correlation to Alameda County (from 0 for no correlation to 1 for an exact correlation), and "Similarity to Active (Rank)," which ranks the counties by degree of correlation.



County	Home Value ($)	Farm Acres	Average Income	Population	+65 Population %	College Graduate %	Unemployment %	Life Expectancy	Similarity to Active	Similarity to Active (rank)
CA Alameda	303100	258070	26680	1443741	10.2	34.9	0.01721361379915	75.3	1	1
OH Lucas	90700	79583	20518	455054	13.1	21.3	0.02653311475122	74.3	0.9995174620639	2
MO St. Louis	63900		16108	348189	13.7	19.1	0.0287343942514		0.9990848639588	3
MN Ramsey	126400		23536	511035	11.6	34.3	0.01261557427573	77	0.9990642545836	4
MA Worcester	146000	103400	22983	750963	13	26.9	0.01654409072085	76.6	0.9988971895812	5
LA East Baton Rouge	98800	66008	19790	412852	9.9	30.8	0.01890265761096	74	0.9988648583875	6
OH Montgomery	95900	105965	21743	559062	13.7	22.8	0.0190622864727	74.9	0.9987963827047	7
NJ Essex	208400		24943	793633	11.9	27.5	0.02649965412225	71.1	0.9986002992664	8
NC Wake	162900	113201	27004	627846	7.4	43.9	0.007493875886762	76.1	0.9985876584171	9
CA Santa Clara	446400	318654	32795	1682585	9.5	40.5	0.01737029630004	77.7	0.9984331036076	10

Figure 7: Alameda County (highlighted) and the 10 most similar counties.

Variations on the Theme

Not all parallel coordinates graphs available in commercial software go by the name parallel coordinates, and they don't all look exactly the same. Besides Spotfire, other business intelligence vendors who offer parallel coordinates graphs include SAS, ProClarity (now owned by Microsoft), Advizor Solutions, and Information Builders (by virtue of the fact that they sell Advizor Solutions' software under a different name through an OEM relationship). To illustrate one more approach to using parallel coordinates, I'll shift over to the product named *Advizor Analyst/X* from Advizor Solutions. Figure 8 provides an example of a parallel coordinates graph (called a *parabox* by Advizor Solutions), which displays multivariate data regarding a company's customers (one line per customer) in a way that looks different than previous examples.

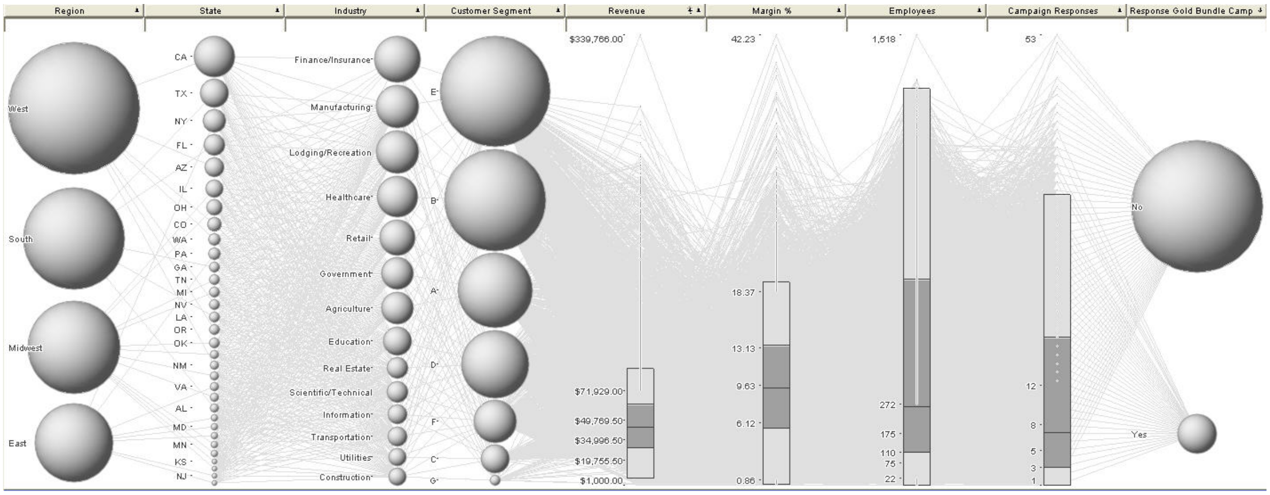


Figure 8: A parabox (another name for parallel coordinates) graph from Advizor Solutions**.**

The variable names appear across the top, including region, state, industry, and so on. This particular example includes both quantitative variables, such as revenue, and categorical variables, such as region. In addition to the gray lines that connect a value of each variable for a given customer, circles display the relative sizes of each value belonging to a particular categorical variable and a box plot displays the distribution of values for a particular quantitative variable. These circles (also known as *bubbles*) and box plots summarize each variable in a way that can't be seen merely by looking at the lines, which is a nice addition (although the 2-D areas of circles cannot be compared precisely).

Parallel coordinates can reveal correlations between multiple variables. This is particularly useful when you want to identify which conditions correlate highly to a particular outcome. For instance, this example can be used to examine which conditions seem to have contributed to the desired outcome of customers responding to a special marketing campaign named the "Gold Bundle Campaign," which appears on the rightmost axis of the graph. As you can see, relatively few customers responded (indicated by "Yes") to the campaign. It would be useful to know the characteristics of those customers who responded. Look at what happens when I select the "Yes" circle on the "Response Gold Bundle Campaign" axis (see Figure 9).

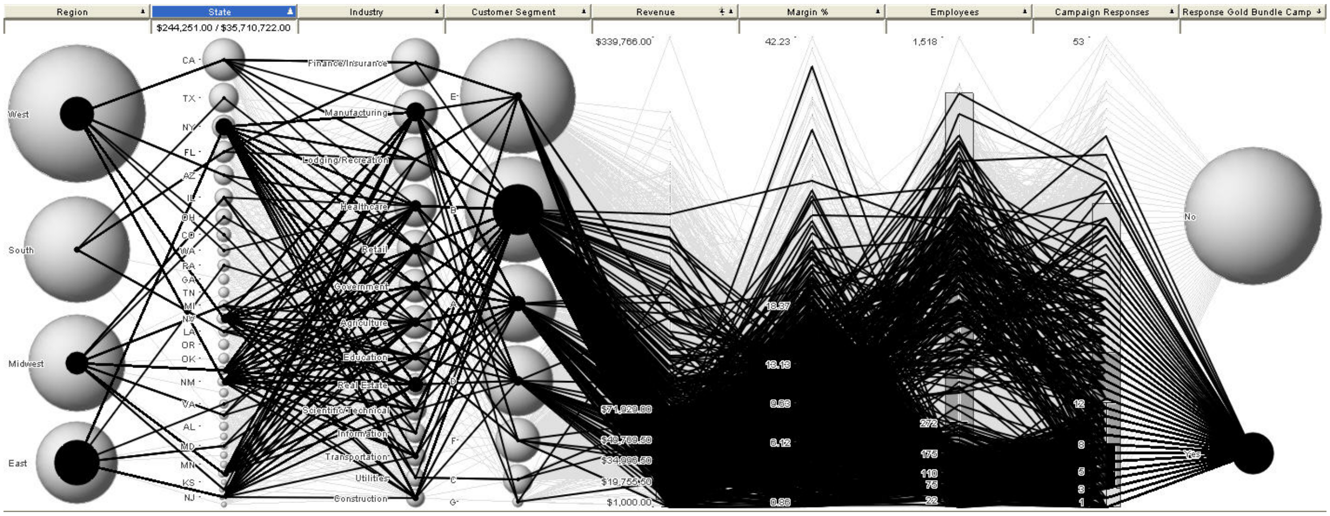


Figure 9: All customers who responded to the Gold Bundle Campaign are highlighted.

Now we can begin to look for predominant characteristics across the other variables. Before we do so, however, I'm going to eliminate some of the clutter by turning off the lines, resulting in the graph that appears in Figure 10.

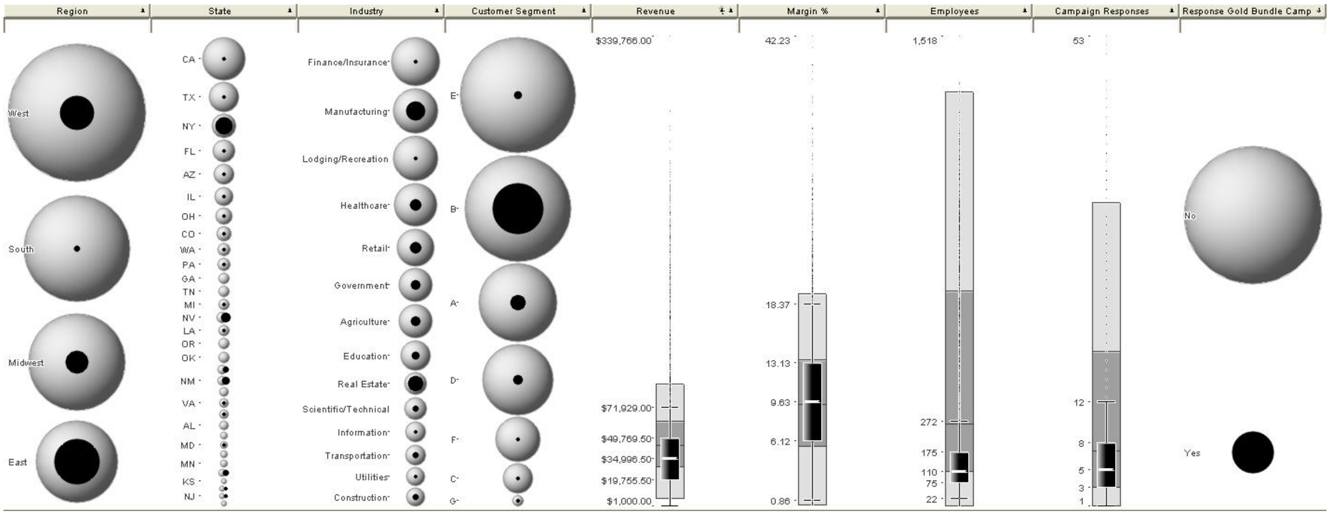


Figure 10: The same display as Figure 9 but without the lines.

Now it's easier to see the relationships. The first thing I notice is that, of the four regions (on the left-most axis), a much greater percentage of customers in the east responded than anywhere else, which appears to be largely a result of a significant response in the state of New York. The industries that responded the most are manufacturing and real estate, with about the same number of responses, but a much higher percentage of real estate customers. Shifting attention to the quantitative variables, I can easily see that responders tended to have lower than average revenues, profit margins that are typical, but a much lower than average number of employees (that is, they are relatively small companies). Another interesting characteristic is the fact that those customers that responded usually respond much less favorably to marketing campaigns, shown on the Campaign Responses axis. This is a good example of what can be discovered when exploring multivariate business data using a well-designed parallel coordinates display.

I hope that you are beginning to get a sense of what can be seen and the useful questions that can be pursued and answered when using parallel coordinates. Multivariate analysis requires specialized visualizations and methods of interaction with data. Parallel coordinates is only one approach. Next month we'll look at what you can do with heatmaps.

About the Author

Stephen Few has worked for over 20 years as an IT innovator, consultant, and teacher. Today, as Principal of the consultancy Perceptual Edge, Stephen focuses on data visualization for analyzing and communicating quantitative business information. He provides training and consulting services, writes the monthly *Visual Business Intelligence Newsletter*, speaks frequently at conferences, and teaches in the MBA program at the University of California, Berkeley. He is the author of two books: *Show Me the Numbers: Designing Tables and Graphs to Enlighten* and *Information Dashboard Design: The Effective Visual Communication of Data*. You can learn more about Stephen's work and access an entire library of articles at www.perceptualedge.com. Between articles, you can read Stephen's thoughts on the industry in his blog.