Investigating Data With Andrews Plots

Neil H. Spencer
University of Hertfordshire
United Kingdom

N.H.Spencer@herts.ac.uk

Most data that are collected are multivariate in nature, and much of them can be regarded as continuous. In the initial stages of analysis, graphic displays can be used to explore the data, but for multivariate data, traditional histograms or two- or three-dimensional scatter plots may miss complex relationships that exist in the data set. A number of methods for graphically displaying multivariate data have been suggested. However, these are not generally available in major statistical packages and are thus largely not used by researchers. One of the most appealing methods is that of Andrews Plots. This article discusses the potential uses of Andrews Plots and makes them accessible to users through the production of a freely available add-in for Microsoft Excel. The use of Andrews Plots is demonstrated by using data from the 2001 Parliamentary General Election in the United Kingdom.

Keywords: multivariate data; Andrews Plots; Microsoft Excel

Introduction

The vast majority of data collection exercises in the social sciences yield multivariate data, and much of these data are continuous in nature or can be approximated as such (e.g., age in years). A first step in any investigation of data is to create simple summaries and graphs so as to obtain a better understanding of the nature of the data before analysis and also check for outliers or unusual observations. However, the production of univariate or bivariate summary statistics (e.g., means, correlations) or of histograms and two-dimensional scatter plots to graphically display the data may hide more complex relationships existing in the multivariate data set.

Methods for the graphic display of multivariate data have been reviewed by Cluff, Burton, and Barrett (1991) and are also discussed in books on multivariate analysis, including Everitt and Dunn (1991). Cluff et al. (1991) looked at Symbolic Star Plots, Chernoff Faces, Glyphs, Boxes, Profile Plots, Symbol Scatter Plots, Kleiner-Hartigan Tree Symbols, Generalised Draftsman Displays, Andrews Plots, Parallel Axes Graphics, Cartesian Hyperspace Graphics, and Kent's Method.

Andrews Plots

The lack of availability in major statistical packages used in social science research (e.g., Statistical Package for the Social Sciences, or SPSS) of the methods of display outlined by Cluff et al (1991) means that they are not, in general, used by researchers. One of the most appealing and visually simple of these methods for comparing cases on the basis of multivariate data is that of Andrews Plots. Introduced in an article in 1972, Andrews Plots (Andrews, 1972) provide a means for the simultaneous display of several continuous variables, using the formula

$$x(t) = x_1 / \sqrt{2} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + x_6 \sin(3t) + x_7 \cos(3t) + \cdots$$

where $-\pi \le t \le \pi$. The technique places greater weight on those variables listed first (e.g., x_1, x_2) and less weight on variables listed later (e.g., x_6, x_7, above). It is thus recommended (e.g., Everitt & Dunn, 1991) that principal components of the original data be produced and used in the order

reflecting their importance, unless there is a natural ordering that can be used. This article brings Andrews Plots into the realm of multivariate techniques accessible to users of statistics through the production of a freely available add-in for Microsoft Excel (please contact author). Through the use of dialog boxes, the user is able to define the variables to be used to produce the plots, and the default (which can be switched off) is to produce Andrews Plots of principal components of the data. Graphs that are then produced (either in colour or monochrome) can be edited as normal Excel graphs by the user if required. When the mouse is pointed at a particular curve of interest, Excel displays the label that the user has identified as case identifiers, enabling simple identification of cases of interest. Facilities are also available in the add-in for producing quantile versions of the plots (Gnanadesikan, 1977) and for plotting only selected cases, identified by a selection variable.

Detecting Unusual Observations with Andrews Plots
Methods for detecting outliers that use data in a multivariate manner have been outlined in, for instance, Krzanowski and Marriott (1994). However, the majority of researchers who wish to use statistics as a tool in their research and do not regard themselves as experts in its use are more likely to prefer graphic methods where interpretation relies less on technical expertise.

Andrews Plots can be used in this regard. A case with an unusual profile of values for the variables used to create the Andrews Plots will have a curve that is, in some way, different from others in the plot. When the Andrews Plots are created using the Microsoft Excel addin, these curves can be identified with particular cases by pointing the mouse at the curve, causing the case identification label to be displayed.

A demonstration of the use of Andrews Plots to detect unusual observations follows later.

Classifying Cases with Andrews Plots
As well as being of great value in detecting outliers, Andrews Plots can be used as a technique for classifying observations into groups.

Where unknown grouping of cases exists within a data set, an Andrews Plot can be used to discover this effect. Cases within one group will have similar patterns of values for the variables used in the Andrews Plot, and similar curves will be produced for these cases. Cases in a second group will have a different profile of values for the variables from those in the first group, and thus the curves produced for this second group will show a different pattern from those for the first group.

Andrews Plots can also be used to allocate a case to one of a number of groups that are known to exist in the data set. Curves for cases whose grouping is known are created and then the curve for the unknown case superimposed. The unknown case is allocated to the group whose set of curves its own curve most closely matches.

Application
To demonstrate the use of Andrews Plots created using the add-in available, data from the 2001 Parliamentary General Election in the United Kingdom are used.

The data used consist of the percentage of the vote achieved by the three main political parties in the United Kingdom: Conservative (Cons), Labour (Lab), and Liberal Democratic (Lib Dem). Constituencies in Northern Ireland have been excluded from the data set as the structure of political parties there does not follow the same pattern as the rest of the United Kingdom. For illustrative purposes, part of the data are shown in Table 1. The three constituencies shown relate to those represented by the leaders of the main parties at the time of the election: William Hague (Cons; Richmond), Tony Blair (Lab; Sedgefield), and Charles Kennedy (Lib Dem; Ross, Skye, and Inverness West).

Table 1

Data From the 2001 Parliamentary General Election
in the United Kingdom (in percentages)

Constituency	Labour Vote	Conservative Vote	Liberal Democratic Vote
Richmond	21.87	58.93	17.92
Ross, Skye, and Inverness West	16.89	8.89	54.10
Sedgefield	65.28	20.99	9.06

Because the three variables in Table 1 do not have a natural ordering, a full set of three principal components was obtained from these data. Using these principal components, Andrews Plots were created using the following formula:
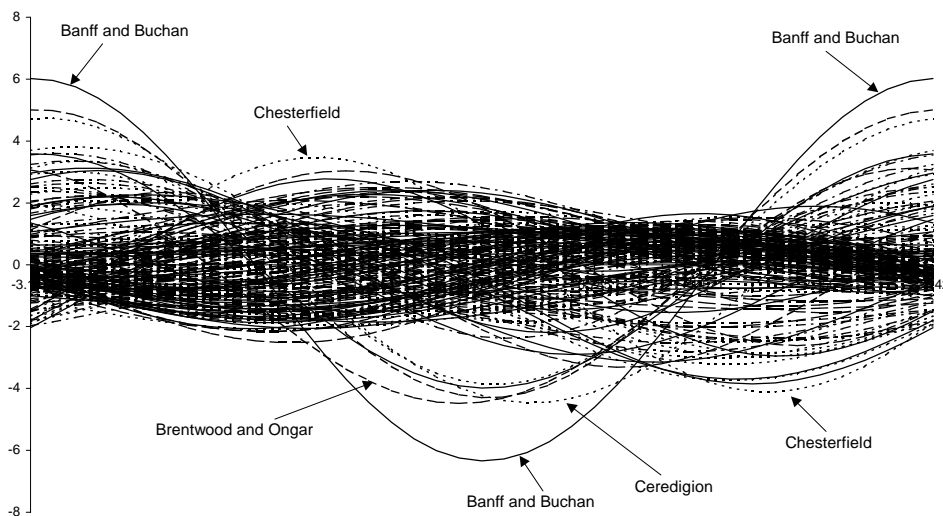
$$x(t) = x_1/\sqrt{2} + x_2 \sin(t) + x_3 \cos(t)$$

where x1, x2, and x3 are the first, second, and third principal components, respectively. This formula is the standard three-variable version of that produced by Andrews (1972) and defines the shapes of the curves on the Andrews Plot. The curves were evaluated for values of t from $-\pi$ to $+\pi$, a total range of 2π. This range was chosen because both sin(t) and cos(t) repeat themselves in such a way that sin(t) = sin(2π + t) = sin(4π + t) = . . . and cos(t) = cos(2π + t) = cos(4π + t) = ... Equivalent curveswould have been obtained for any set of values of t that had a range of 2π.

The curves are displayed in Figures 1, 2, and 3. The horizontal axis represents t and the vertical axis represents x(t). Because Microsoft Excel cannot display more than 255 curves per chart, the add-in has spread the cases across three displays as evenly as possible. What is seen in Figures 1 through to 3 is the result of allowing the default settings to operate. The addin does have the facility to allowthe user to choose which particular cases to display in a plot, if required. The labels on the figures are the constituencies associated with the curves. They were identified by pointing the mouse at the curve, causing Microsoft Excel to display the case identifier (constituency name).
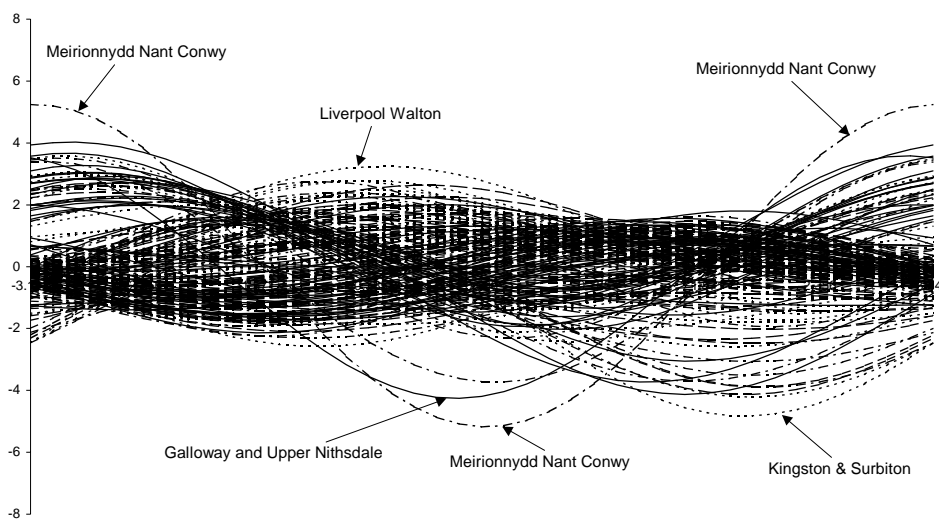
Figure 1

Andrews Plots: Cases 1 to 214

3

In Figure 1, one curve that signifies an unusual observation is that for the constituency "Banff and Buchan" where all three major parties polled very poorly as it is the seat of the former Scottish National Party leader, Alex Salmond, who won with 54.2% of the vote— their highest percentage in a constituency. Another curve of note here is for the constituency "Ceredigion" where Plaid Cymru (the Welsh Nationalist Party) won but, unusually, the Liberal Democrats also polled well. A third curve that warrants mention is that for Chesterfield. This was formerly the constituency of a famous Labour politician, Tony Benn, but in 2001, the Liberal Democrats won the seat with 47.8% of the vote, with Labour in second with 42.0% and the Conservatives trailing a long way behind with 8.1%. "Brentwood and Ongar" is unusual in that an independent candidate, Martin Bell, stood against the sitting Conservative Member of Parliament and finished in second place, ahead of both Labour and the Liberal Democrats.

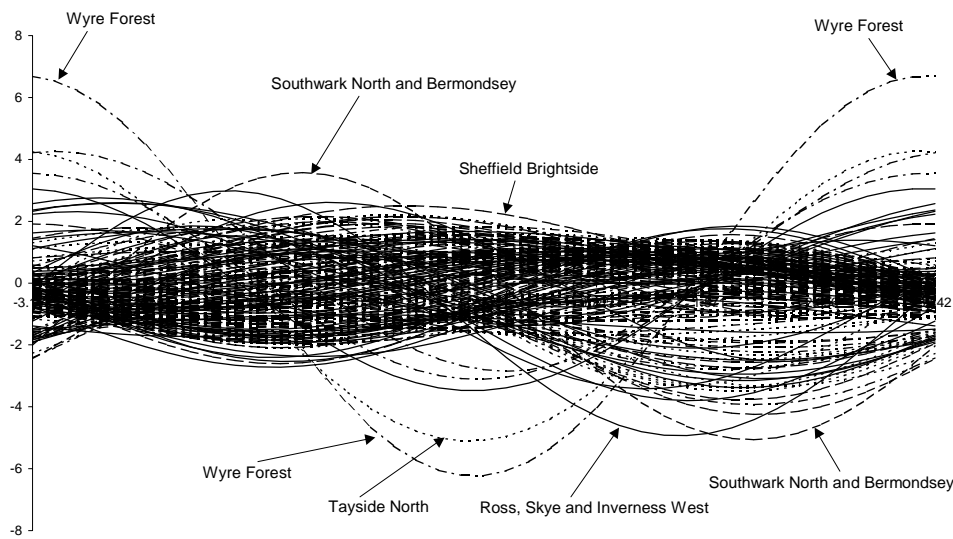## Figure 2

Andrews Plots: Cases 215 to 428

The clearly most unusual curve in Figure 2 is that for "Meirionnydd Nant Conwy," which was the constituency won by the leader of Plaid Cymru, Elfyn Llwyd, and gave his party its largest victory. Similarly, "Liverpool Walton" is notable because it is the constituency that gave Labour its highest percentage of votes, and "Kingston & Surbiton" is notable because it is the constituency that gave the Liberal Democrats its highest percentage of votes. Unusually in the "Galloway and Upper Nithsdale" constituency, the Conservative and Scottish Nationalist parties obtained 67.8% of the vote between them, with Labour and the Liberal Democrats trailing.

There are a number of curves of note in Figure 3. "Wyre Forest" is a constituency where an independent candidate (with a campaign related to the local hospital in Kidderminster) won with a large majority and the Liberal Democrats did not field a candidate. "Southwark North and Bermondsey" and "Ross, Skye, and Inverness West" are constituencies held by the well known Liberal Democrats Simon Hughes and party leader Charles Kennedy. "Tayside North" is the constituency where the Scottish Nationalist Party recorded its second highest percentage of the vote. "Sheffield Brightside" is the constituency represented by David Blunkett, the current Home Secretary who obtained one of Labour's largest votes, in percentage terms.

4

Figure 3

Andrews Plots: Cases 429 to 641

Discussion

This article demonstrates that Andrews Plots can be beneficial to an investigation of social science data and provides an easy-to-use Microsoft Excel add-in to facilitate their creation. Some uses of Andrews Plots have been discussed, and an application to election data demonstrated to highlight unusual cases.

It is hoped that researchers working in the social sciences will recognize the potential usefulness of Andrews Plots and make use of the add-in to gain a greater understanding of their data.

References

Andrews, D. F. (1972). Plots of high dimensional data. Biometrics, 28, 125-136.

Cluff, E., Burton, R. P., & Barrett, W. A. (1991). A survey and characterization of multidimensional presentation techniques. Journal of Imaging Technology, 17(4), 142-153.

Everitt, B. S., & Dunn, G. (1991). Applied Multivariate Data Analysis. London: Edward Arnold.

Gnanadesikan, R. (1977). Methods for statistical data analysis of multivariate observations. NewYork: JohnWiley.

Krzanowski,W. J., & Marriott, F. H. C. (1994). Multivariate analysis: Part 1, Distributions, ordination and inference. London: Edward Arnold.