A Remarkable Scatterplot

Ian Spence; Robert F. Garrison

The American Statistician, Volume 47, Issue 1 (Feb., 1993), 12-19.

A Remarkable Scatterplot

IAN SPENCE and ROBERT F. GARRISON*

The Hertzsprung-Russell diagram, invented by astronomers just after the turn of the century, has had a profound influence on almost every aspect of modern astronomy, but especially on theories of stellar evolution. The diagram is an excellent pedagogical example of a scatterplot and is remarkable on many counts: it is an impressive illustration of the power of graphical display; it may be used to show the strengths and limitations of smoothing, or the fitting of parametric functions; and variations of the plot can be used to represent change over time, large numbers of data points, or differing relationships in distinct strata.

KEY WORDS: Data analysis; Modeling; Graphical perception; Astrophysics; Stellar evolution; Hertzsprung-Russell diagram; History of science.

1. INTRODUCTION

Every good teacher of statistics advocates the use of graphs, often paraphrasing Tukey's (1977) declaration: "**The greatest value of a picture** is when it *forces* us to notice **what we never expected to see**" (emphasis and italics in original). In isolation, this sound advice may be less than convincing to beginning students who have little experience of either statistical graphs or data. The purpose of this article, therefore, is to provide a pedagogical illustration using a graph that profoundly influenced thinking in astronomy. The subject matter is esoteric, but of general interest, and the science is important, yet easy to present in simple terms.

In the physical, biological, and social sciences, the predominant graph is the scatterplot, appearing in many variations (see Cleveland and McGill 1984 for examples); indeed, it has been estimated that 75% of the graphs used in science are scatterplots (Tufte 1983). Data analysts often use simple scatterplots before trying other forms of analysis, and the insights gained may stimulate the production of more complicated variations or may guide the choice of a model. Linear or nonlinear relations are easy to discern and the human eye is robust to the effects of outliers and other aberrations in the data (Mosteller, Siegel, Trapido, and Youtz 1981; Spence and Lewandowsky 1990; Wainer and Thissen 1979).

Perhaps the most spectacularly successful example of a simple scatterplot was first shown during a lecture given by the American astronomer Henry Norris Russell in 1912 (Russell 1913). This diagram profoundly affected the way in which astrophysicists think about stellar evolution and it continues to do so. It has been described as "one of the greatest observational syntheses in astronomy and astrophysics" (Smith and Jacobs 1973) and versions of it appear in every introductory textbook of astronomy. In addition, variations and elaborations of Russell's original chart are to be found throughout the technical literature of astronomy (Davis Philip and Hayes 1978; Garrison 1991).

The diagram is a plot of the brightness of stars versus their spectral class. The former is given in units of *absolute* magnitude, a measure of light energy *corrected for distance*. *Apparent* magnitude is brightness *uncorrected* for distance and is therefore unsuitable if we wish to make comparisons among stars. Spectral class is established by the appearance of the star's spectrum. Although the classes were originally ordered alphabetically, according to the strengths of the hydrogen lines, it soon became apparent that temperature is a more fundamental parameter and that the maximum of the hydrogen line strength does not occur at the maximum temperature. Thus, the conventional sequence has the peculiar order OBAFGKM, with O-type stars being the hottest and M-type stars the coolest; some intermediate letters have been dropped as redundant. Subdivisions labeled 0–9 are commonly used to refine the classes. Spectral class is also closely related to color and surface temperature; the hottest stars emit most of their light in the blue region of the spectrum, and the coolest stars appear red.

In the system used by astronomers today, stellar magnitudes range from negative values for bright stars (for example, Sirius, the brightest star seen, has *apparent* magnitude -1.5) to large positive values for very faint objects (stars that we can just see have *apparent* magnitudes of about $+6$). This peculiar system dates from the second century B.C., when Hipparchus compiled a star catalog, referring to the brightest stars as being of the *first magnitude* and the faintest as being of the *sixth magnitude*. Both Hipparchus's and the more refined modern scale are logarithmic, in approximate correspondence to the eye's responsiveness to brightness: 5 magnitudes correspond to a factor of 100 in brightness and 1 magnitude to a factor of 2.5.

The Hertzsprung-Russell, or HR, diagram is important because the observable variables, absolute magnitude and spectral class, are closely related to the fundamental variables of mass, age, and chemical composition, which in turn determine the evolution and structure of a star.

2. HERTZSPRUNG-RUSSELL DIAGRAM

Representations of data on the plane are very old indeed. One example, a map of Northern Mesopotamia

---

*Ian Spence is Professor, Department of Psychology, and Robert F. Garrison is Professor, Department of Astronomy, University of Toronto, Toronto, Ontario M5S 1A1, Canada. This work was supported by the Natural Sciences and Engineering Research Council of Canada: Grant A8351 to Spence and Grant A5478 to Garrison. The authors thank Linda Tilley for help in preparing the figures. Nancy Houk kindly provided us with an original copy of Figure 8 and permission to reproduce it.

on a clay tablet, dates from 3800 B.C. (Beninger and Robyn 1978.) Notwithstanding, a mathematical formalization of the coordinate system implicit in the construction of maps was absent until Descartes's *La Geométrie* of 1637. Yet it was more than a century thereafter that Cartesian data plots with superimposed hand-drawn functions first appeared (Lambert 1779), and true scatterplots do not emerge until the late 19th century. Thus, despite its relatively recent invention, Russell's scatterplot, which first appeared in print in 1914, is an early instance of the form.

When two variables are linearly associated, it is not too hard to detect the relation in a table of numbers, but when the function is more complicated, as for instance when X maps into two or more Y's, it is extremely difficult to discover the structure by examination of the table alone, even when the data are ordered with respect to one of the variables. The Danish astronomer Ejnar Hertzsprung collected and discussed data that were difficult to visualize (Hertzsprung 1905, 1907), and the problem was compounded because the data were not assembled in a single table, but in a collection of small tables (the result of work by several astronomers). It seems that Hertzsprung constructed a variety of charts in his efforts to understand these data, but no graph appears in either the 1905 or the 1907 paper (however, a diagram of apparent magnitude versus color index for stars in the Pleiades and Hyades clusters may be found in the Potsdam Astrophysical Observatory Publications for 1911.)

Henry Norris Russell had access to these and similar data and his inspiration was to put them together in a single graph relating absolute magnitude to spectral type. His diagram was shown publicly during an address to the Royal Astronomical Society in 1912, and subsequently mentioned in print (Russell 1913, p. 324), but not reproduced. A later paper (Russell 1914), based on a talk given within six months of the original address, includes a diagram that is probably similar, if not identical, to the "slide shown on screen" in 1912. The graph is Figure 1 in Russell (1914) and Figure 1 in the present article. As in our reproduction, the axes were not labeled and the two sloping lines were present in the original. The large circles in the graph represent groups of stars and other variations in the smaller symbols are related to the trustworthiness of the observations, but are not significant for the purpose of this article.

The data are not distributed at random, showing all possible combinations of brightness and temperature: rather, they are found principally in certain regions of the scatterplot. Russell's original description (1914, p. 287) summarizes the essentials:

> The appearance of Figure 1 therefore suggests the hypothesis that, if we could put on it some thousands of stars, instead of the 300 now available, and plot their absolute magnitudes without uncertainty arising from observational error, we would find the points representing them clustered principally close to two lines, one descending sharply along the diagonal, from B to M, the other starting also at B, but running almost horizontally.

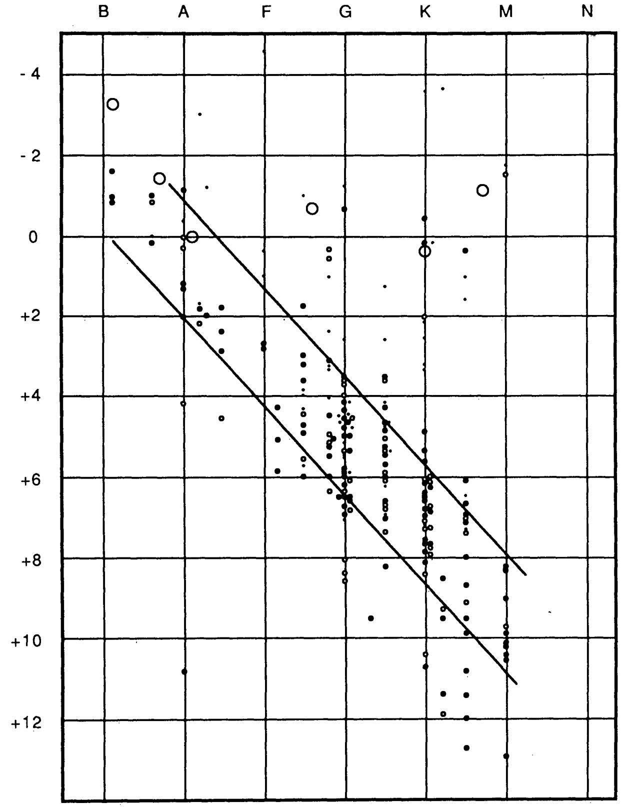

Figure 1. Russell's Figure 1, a Plot of Absolute Magnitude Against Spectral Class. Since the original does not reproduce well, this is a scanned and edited version, but it is substantially identical to the original.

Later on the same page, he continues,

> We may express this hypothesis in another form by saying that there are two great classes of stars—the one of great brightness, (averaging perhaps a hundred times as bright as the Sun) and varying very little in brightness from one class of spectrum to another; the other of smaller brightness, which falls off very rapidly with increasing redness. These two classes of stars were first noticed by Hertzsprung, who has applied to them the excellent names of *giant* and *dwarf* stars.

A modern version of the diagram, as seen in elementary textbooks, is shown in Figure 2. Pedagogical versions often show the inferred surface temperatures, in addition to the spectral class and, in general, they do not represent actual data, but tend to be schematic.

3. LIFE OF A STAR

The Hertzsprung-Russell diagram has had a tremendous effect on the development of astrophysics, especially theories of stellar evolution. To appreciate its importance, we must have some understanding of how stars evolve. The lifetime of a typical star like the Sun is on the order of 10 billion years, so we cannot see stars changing, except for rare and sudden events such as supernovas. Part of our conception of how stars are born, develop, and ultimately die, comes from terrestrial physical theory, and part comes from studying the "snapshots" provided by Hertzsprung-Russell diagrams, where we observe many stars, each at a different stage in its evolutionary cycle. Position in the diagram is determined by the star's mass, chemical composition, and age, with the mass remaining more or less constant

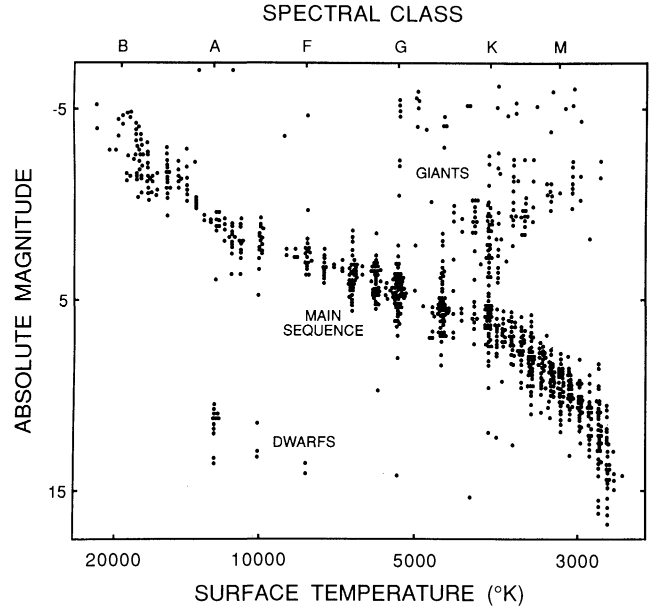

Figure 2. A Modern Hertzsprung-Russell Diagram, Using Actual Data.

during the life of a star. However, the star moves in the diagram as it ages. Figure 3 summarizes the hypothesized evolutionary track of a star, about the same size as our own Sun; the ages shown are approximate. The tracks of stars with larger or smaller masses differ in shape and position and stars with larger masses evolve much faster than smaller ones.

Stars begin life as a cloud of interstellar gas and dust. When the cloud reaches a critical density, it begins to contract gravitationally and a large cloud may break up into many fragments, forming a cluster. When the central temperature and density of a fragment are high enough, nuclear fusion of hydrogen into helium takes place and a self-luminous star is born, with the resultant outflow of energy balancing the contraction. Because the gas making up the star is mostly hydrogen, which is the primary source of energy, the star remains in this stable balance for most of its life cycle. The location of a star on the main sequence depends on its mass; thus the main sequence is essentially a mass sequence, with the smallest masses in the lower right corner and the largest masses in the upper left. Stars spend 90% of their active lifetimes in this stable state and hence the hydrogen burning main sequence is the locus of points representing most known stars.

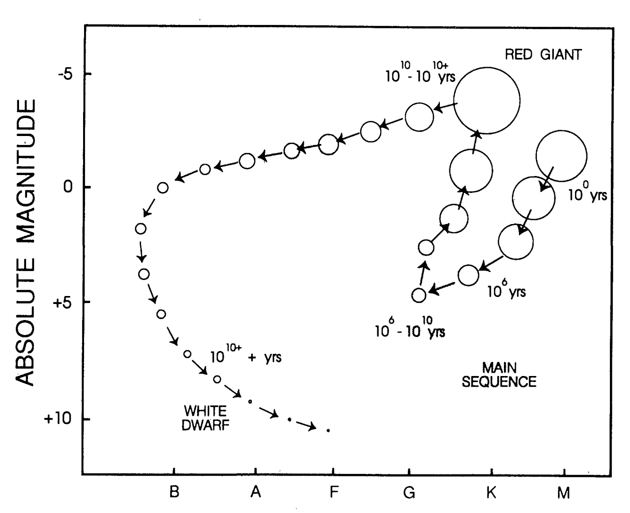

Figure 3. A Schematic Hertzsprung-Russell Diagram Showing the Developmental Sequence for a Single Star Like Our Sun. The approximate age is shown at various stages. About 90% of the star's active life is spent on the main sequence.

Stars exhaust the hydrogen fuel in the central region and must adjust their interior to allow helium to fuse into carbon, nitrogen, and oxygen at a rate which depends on mass. A star like the Sun is stable on the main sequence for about 10 billion years. Very massive stars exhaust their hydrogen much more rapidly than the Sun (hundreds of thousands of years only), whereas low-mass stars may stay on the main sequence for a hundred billion years.

Helium fusion releases more energy, so the delicate balance between gravitation and energy outflow, which characterizes the main sequence stars, is upset. Stars in the helium fusion stage become much larger and their outer layers become cooler and less dense. The atmosphere of such a red giant would fill the Earth's orbit! As they age, stars of many different masses funnel into the region of the red giants (near spectral class K and absolute magnitude 0.) This accounts for the concentration of stars in that region in spite of the fact that an individual star spends less than 10% of its lifetime there.

Finally, all available fuel exhausted, most dying stars collapse to the size of the Earth. This is the white dwarf phase, when a balance is achieved between gravitation and the pressure of crowded atomic nuclei, with subsequent slow cooling to a black cinder. No further fusion is possible. Only very massive stars experience the spectacular death of a supernova explosion, which results in either a neutron star (about 20 km in diameter) with a balance between gravitation and neutron density, or a black hole (less than a few kilometers in diameter) with no well-understood balance.

All this, and much more, has been deduced with the aid of the HR diagram.

4. STATISTICAL ANALYSIS

Scientists routinely use a variety of statistical techniques to assist in detecting structure in data. In this section, we consider the application of regression and clustering procedures to Russell's data to highlight some strengths and weaknesses of such methods. These illustrations are not intended to be exhaustive but to emphasize that good data analysis cannot be mechanical but must always be attentive to the unexpected. The following analyses are presented to exemplify the approach of a typical working scientist with a knowledge of basic statistics.

Russell (1914) did not present his data in tabular form. The data set must be approximated by digitizing the points shown in his Figure 1, and it is necessary to make some assumptions when doing this. For example, the original graph contains six large circles which Russell (1914, p. 286) says "represent mean results for numerous bright stars." At one point, he states (p. 287) that 300 stars were used. Since there are 220 single points in his diagram, this implies that the large circles represent 80 stars in total. However, elsewhere (p. 286) he says that the circles represent about 120 stars, which would yield a total of about 340 stars, not 300. We have assumed that each of the six circles represents 14 stars, identical in brightness and temperature; thus our version of Russell's graph contains 304 stars. The exact number of stars was not crucial to Russell's argument, however, nor is it for our purposes.

We adopt another assumption implicit in Russell's treatment of the data. The position of a star with respect to the abscissa is determined by its spectral class. Russell chose to space these classes equally along the axis and we have done likewise, assuming equal intervals among the categories. Since spectral class is a surrogate for both color and surface temperature, this is not unreasonable, although other assignments of scale values to the classes are possible—a topic beyond the scope of this article.

It is interesting to see what ordinary linear regression produces if brightness is regressed on spectral class. The regression is shown in Figure 4 and the associated R^2 is .31 (Pearson correlation = $-.56$). The line lies at an angle to the main sequence and has clearly been strongly influenced by the giants. Although numerical indices such as R^2 and tests of significance would suggest that there is a substantial linear association between brightness and temperature, it is probably fair to say that linear regression has failed to provide a good summary of the data. Paradoxically, the fitting of a *single* line has made it more difficult to see the better description using *two* lines: one aligned with the main sequence, and the other approximately horizontal through the giants, as suggested by Russell (1914).

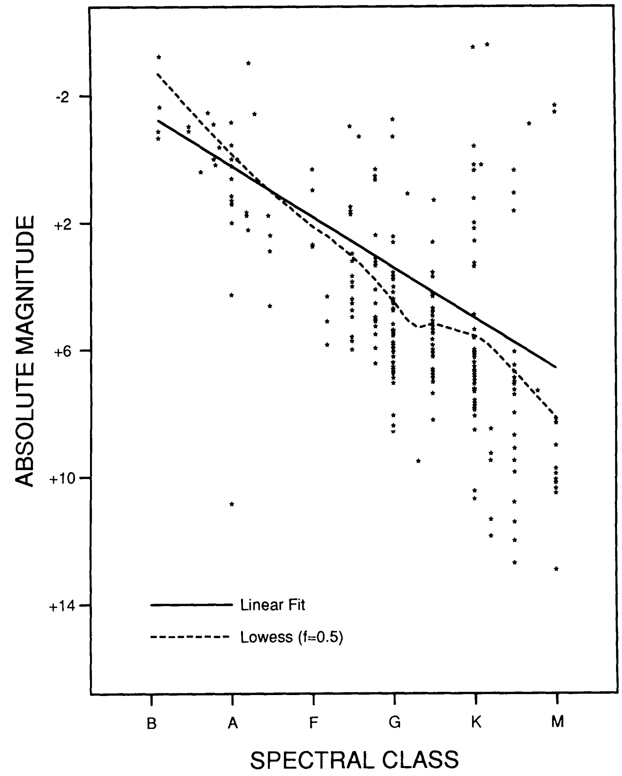

Figure 4. The Least Squares Linear Fit and Lowess Smoothed Fit to Russell's Data. The giants exert considerable leverage on both fitted functions.

An alternative to conventional regression is the use of a "smoother." There are several possibilities, but the one we have used is called *lowess*—locally weighted scatterplot smoothing, a robust fitting procedure described by Cleveland (1979). The broken line in Figure 4 results when the lowess technique is used with a smoothing parameter of $f = .5$. The New S statistical system (Becker, Chambers, and Wilks 1988) provides a convenient way of fitting and plotting lowess functions. Other values of the smoothing parameter f yield similar results, but with a greater or lesser degree of jaggedness. The smoothed function comes closer to alignment with the main sequence than the linear regression, but it is still adversely influenced by the giants.

Model fitting is not complete when we find a function that roughly describes some or even most of the data. Data points that lie far from the fitted function may be even more important than those that are well approximated by the function. These seemingly inconvenient departures may be the most interesting aspect of the fit. In Russell's diagram, the very fact that there are observations that cannot adequately be explained by any function oriented from top left to bottom right cries out for further investigation and explanation—all the more so because it is known that observational errors are negligible. Although the description of data by fitting functions is an important part of data analysis, accounting for deviations from these functions is just as important. Residuals must not be ignored.

A common tactic is to plot the residuals, or standardized residuals, against the fitted values (Draper and Smith, 1981; Weisberg 1980), and this has been done for the linear regression shown in Figure 4 (see Fig. 5 for the residual plot). In the plot, the natural orientation of axes has been retained—bright stars with negative magnitudes at top and red stars at right, even though this means that the residuals in the top half of the plot have negative sign. Three features are immediately apparent.

First, there is the strong visual impression that there are somewhat fewer residuals above the reference point of zero (corresponding to the brighter stars) than below. This is principally an artifact of overplotting: some points represent several stars that Russell lumped together and summarized by circles. Perceptual distortions of this kind can be pronounced and we discuss some approaches to the problem in Section 5.

Second, the residuals fan out from left to right. Beginning students of regression may feel the urge to transform the absolute magnitudes in an attempt to stabilize what looks like heterogeneity of variance—increasing variance with decreasing brightness. This is

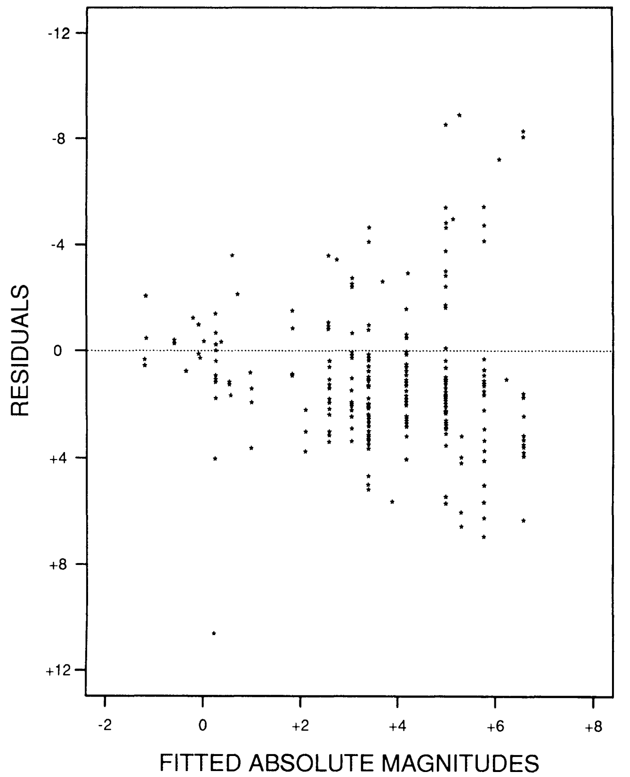

Figure 5. Residual Plot for the Linear Fit to Russell's Data. Brightness decreases to the right since negative values represent the brightest stars.

because they have been taught to look for patterns very similar to this, and for which the normally recommended cure is variance stabilization via a one-bend transformation. Alas, here it is the wrong remedy. The fanning out is a consequence of the fact that two separate relations are present and neither has been accounted for by the single fitted line. As Russell observed, it is possible to draw two lines that yield well-behaved residuals. The residual plot in Figure 5 strongly suggests that the linear fit is unsatisfactory, but its diagnostic value is *not* to tell us to transform the data.

There is one further obvious feature, namely the lone outlier at bottom left, a very faint star whose color is much bluer than the other very faint stars. This aberration plainly deserves attention: is it an error of observation or recording, or is it a truly unusual star? Russell was unsure, noting that the spectrum of this star was difficult to measure because of its proximity to a far brighter binary companion (Russell 1914, p. 286). With the benefit of more recent, accurate measurements, we now know that this star, 40 Eridani B, is a "white dwarf," a collapsed star, at the end of its life cycle, about the size of the Earth, but with the mass of the Sun.

Although the identification of aberrations receives considerable attention in most modern statistical courses, the emphasis sometimes seems to be on disposing of embarrassing data by searching for sources of technical error or minimizing the influence of inconvenient data by the application of resistant methods. Working scientists often find the most interesting aspect of the analysis inheres in the lack of fit rather than the fit itself. The main trend of the data may be of minor significance relative to the importance of the aberrations. Data that are not well accounted for by the model are often the spur to further investigation, modification to theory, or improvement in instrumentation. In the HR diagram, the role of outliers and deviations from the main sequence has been particularly important—in superficial applications of statistical model fitting, these might have been discounted. Indeed, some early workers either ignored the "white dwarfs," located in the lower left corner, or questioned the validity of the distance data (if they were further away, they argued, the absolute magnitudes would be higher, placing them on the main sequence). They are now known to represent the final stages in the life of a star and provide extremely important data for understanding stellar evolution. Thus, the wise consulting statistician will always keep an open mind regarding the importance of data that refuse to fit the mold; not all outliers are bad and robust fitting procedures are not a panacea.

Clustering procedures are a useful adjunct to regression. The literature on clustering and classification contains several techniques (Sneath and Sokal 1973) that can help to discover structure in data. We present one procedure that is sometimes used to reveal clusters and links in data that may be represented on the plane. The minimum spanning tree is the set of straight lines, or edges, connecting all points such that there is a path between any pair of points and the total distance is the minimum possible. There must be N-1 edges and no cycles or circuits. Using New S (Becker et al. 1988), the minimum spanning tree was fit to Russell's data, revealing the main sequence (see Fig. 6) as a long snak-

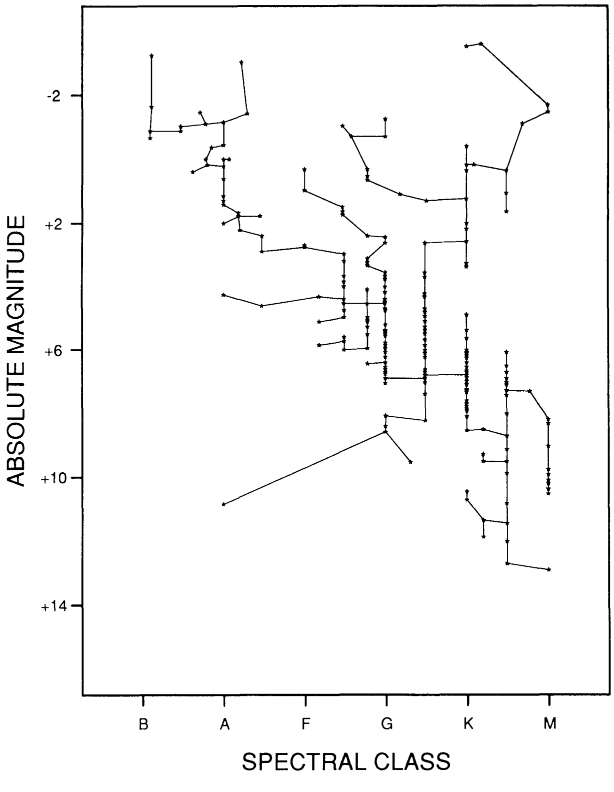

Figure 6. The Minimum Spanning Tree Fit to Russell's Data.

ing path that traverses the set of stars from upper left to lower right. There are two significant branches off the main sequence; they point to the red giants and the solitary star of spectral class A. Hence, the major branches of the minimum spanning tree coincide loosely with the principal features of a Hertzsprung-Russell diagram, namely, the main sequence and the red giants. While no parametric function or model is fit, the structure of the tree may provide useful insights regarding the structure of bivariate or multivariate data (Friedman and Rafsky 1981).

5. PLOTTING LARGE NUMBERS OF STARS

Although Russell's original plot contained only a few hundred stars, data for hundreds of thousands of stars now exist and it is possible to construct Hertzsprung-Russell diagrams in which, because of overlapping points, a misleading impression of relative density in different regions of the plot might be given. Even with some 300 points, Russell (1914, Fig. 1) had to cope with the problem. He plotted points with almost identical measurements next to each other, instead of overlapping, so that their multiplicity would be evident to the reader, and he also used large circles to represent the data from six groups of stars with similar characteristics. Overplotting can be a problem where data points number in the hundreds, or thousands, but distortions can also arise with small numbers of points if either variable is discrete, as in Russell's case where there is a limited number of spectral classes, or if the plotting device lacks sufficient resolution. In such circumstances, it may be preferable to use some form of representation other than plotting each individual point. We show one approach devised by Cleveland and McGill (1984) and another credited to the astronomers Houk and Fesen (1978).

Cleveland and McGill (1984) proposed a device known as *sunflowers* to accommodate large numbers of data points. Figure 7 shows Russell's data plotted using sunflowers. Where two or more points must be plotted at approximately the same location, a sunflower is drawn. This is a set of short lines of equal length, radiating from a single point. The scatterplot is arbitrarily subdivided into many small rectangular regions and the number of lines in each sunflower is equal to the number of corresponding data points. When there are many radiating lines, the resulting icon looks something like a sunflower. Some particularly good examples are to be seen near the top of Figure 7. When there are only two or three data values in the same vicinity, the number of radiating lines is smaller and the plotted symbol is less flower-like. As Figure 7 shows, the relative density in regions with many stars is more faithfully portrayed than if overplotting were allowed (see Fig. 4, for example). Also, it may be argued that it is much easier to see the two great star groupings in Figure 7 than in Figures 1 and 4. However, this example suggests that sunflowers are probably useful only when there are relatively few points, ranging from the hundreds to at most the low thousands. When the number of points is large, sunflowers degenerate into undifferentiated filled circles, even though they represent differing numbers of stars and the scatterplot can look very cluttered.

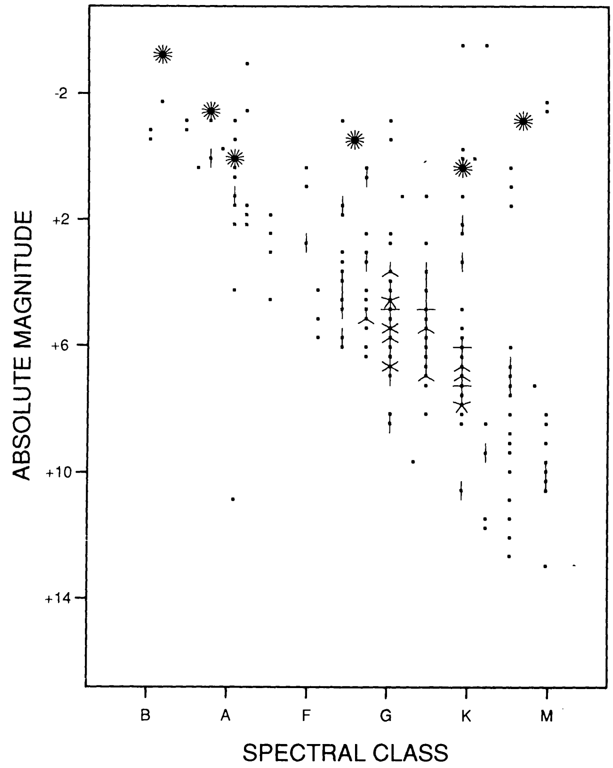

Figure 7. Sunflower Diagram Showing Relative Density in Russell's Data.

Houk and Fesen (1978) plotted data for tens of thousands of stars, substituting octagons of varying size to summarize the data for large groups of stars. The distances between the opposing sides of the octagons are proportional to the number of stars represented (see Fig. 8 for a plot with 93,000 stars) and the impression of differing density in various areas of the plot is compelling. One possible drawback, however, is the use of distance between the sides, rather than area, which may give a misleading impression of greater density in some regions and lesser density in others. Other geometrical forms could also have been used—squares and circles come immediately to mind, although the latter are perhaps to be preferred since there is minimal occlusion in the event of overlap (Cleveland and McGill, 1984).

The Houk plot reveals a mystery that has not yet been resolved: why is there a gap on the main sequence in the region of coordinates (A, 2.0)? The low density of stars at this point on the main sequence does not appear to be an artifact of the scales chosen for the axes, or the plotting technique, but no completely satisfactory explanation has yet been found. Another interesting feature is the cluster of stars in the red giant region near coordinates (K, 1.0). It is known that this grouping, which contains several thousand stars, is not homogeneous with respect to mass—some of these stars are two to three times as massive as others—whereas the main sequence is a "mass sequence," with massive stars at upper left and low mass stars at lower right.

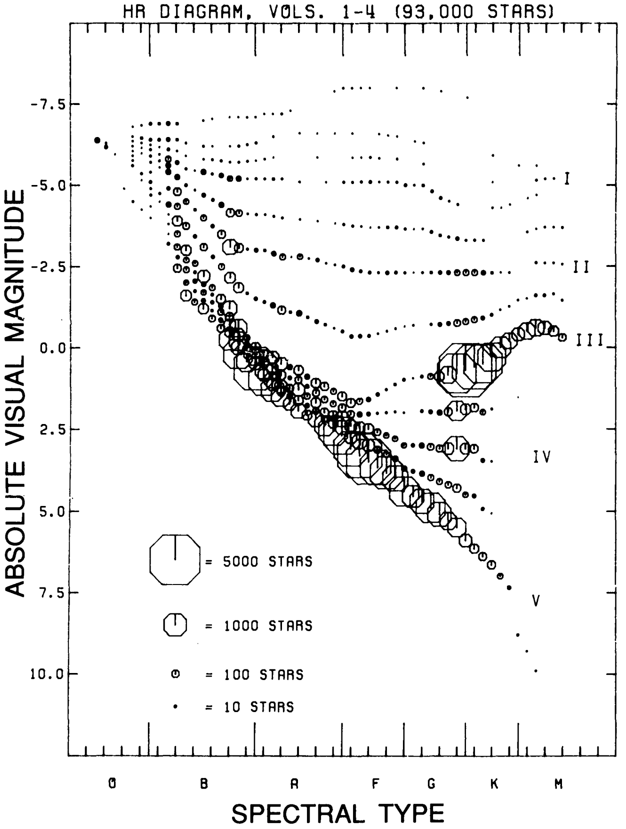

Figure 8. Octagon Diagram Showing Relative Density for 93,000 Stars. The Roman numerals denote the supergiants (I), giants (III), and main sequence dwarfs (V). Courtesy of Dr. Nancy Houk, University of Michigan.

Since stars of differing masses funnel into this area after leaving the main sequence, parameters other than temperature and luminosity must be used to distinguish stars of different mass in the red giant region, whereas on the main sequence, temperature and luminosity suffice to distinguish relative masses uniquely.

Almost a century after it was first devised, the Hertzsprung-Russell diagram continues to stimulate new directions of enquiry in astronomy (Garrison 1988, 1991).

6. DISTINCT STRATA

In Baade's 1944 paper, a seminal work on galactic structure, two distinct strata are plotted in a single Hertzsprung-Russell diagram. To differentiate between the stellar population of the solar neighborhood, which is relatively young, and that of an old globular cluster of stars, Baade used shading for one group of stars and hatching for the other, whereas our version of his diagram (Fig. 9) uses the letters "N" and "G" for the stars in each cluster. Baade's diagram stimulated work on the concept of stellar populations and on the structure of our Milky Way Galaxy. He applied the ideas to galaxies in general and the concept of stellar populations became an important tool for assessing the structure and evolution of galaxies.

The importance of Baade's contribution may be seen more clearly in a later elaboration of his diagram (Sandage 1957; see Fig. 10). Sandage superimposed the HR diagrams of several clusters, using broad lines to represent the distribution of stars in each cluster. The eye quickly discerns that the points of departure from the main sequence are different. Because the main sequence is a mass sequence, this means that the average mass varies from cluster to cluster. Using this insight, the astronomer's task is to determine what other variables might affect the mass. From additional information, such as the location of the clusters in the Galaxy and models of stellar evolution, which show that massive stars (upper left) evolve faster and spend less time on the main sequence than do low-mass stars, the conclusion is drawn that the age of cluster determines the point of departure from the main sequence. Once calibrated, this provides astronomers with a very powerful tool for determining the age of a star cluster, that is,

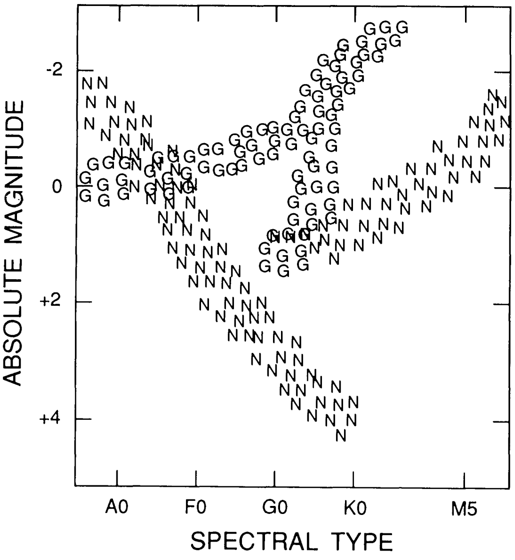

Figure 9. HR Diagram Showing Two Distinct Strata of Nearby and Globular Cluster Stars. After Baade (1944).

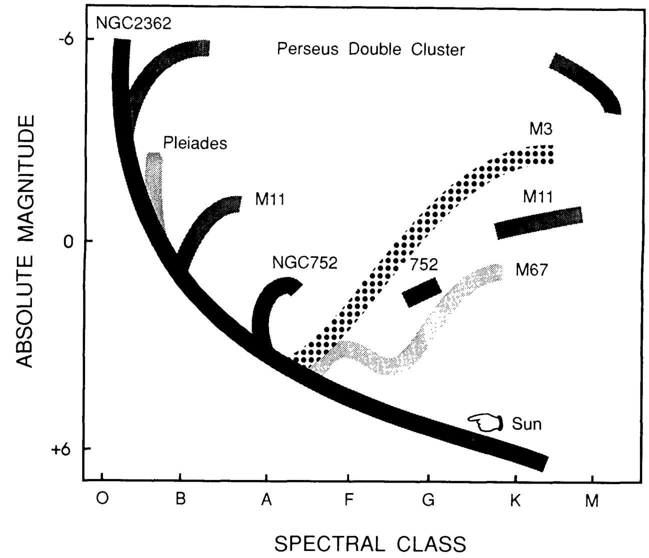

Figure 10. Schematic HR Diagram Showing Several Galactic Clusters and One Globular Cluster (M3). After Sandage (1957).

observing the turnoff point in the HR diagram. Also, the shape of the giant branch differs for clusters with different composition. For example, M3 and M67 are not much different in age, but they have notably different shapes in the HR diagram. This was an important discovery (and a direct result of the use of the HR diagram) because it prompted an examination of the differences in chemical composition associated with age, location, and the general buildup of heavy elements in the Milky Way Galaxy.

Similar devices for encoding multiple strata are frequently seen in published scatterplots in other areas of empirical research. Lewandowsky and Spence (1989) investigated the effectiveness of several coding schemes and concluded that the use of color is optimal when both speed and accuracy of perceptual discrimination are considered. When color is not available, Lewandowsky and Spence (1989) found that *discriminable* letters are effective, and these can also serve a direct mnemonic function. We have redrawn Baade's (1944) diagram with the letter N denoting nearby stars, and the letter G denoting globular cluster stars (Fig. 9). These two letters were chosen because they are highly discriminable; if confusable letters like H, E, or F were used, the reader would have more difficulty in distinguishing the strata. The table of perceptual features in Lewandowsky and Spence (1989) notes that the two letters N and G share no major perceptual features, and so are easily discriminated.

7. SUMMARY

The Hertzsprung-Russell diagram is a shining example of the power of graphic display. Although the graph is simple, it is far from trivial, and astronomers have exploited and extended this remarkable scatterplot in countless ways. There are so many billions of stars that astronomers need a filter like the HR diagram to help them pick out the interesting objects. Often peculiar stars—outliers in the diagram—give valuable insight into the behavior of normal stars. Their location in the diagram gives easy evidence to the astronomer that they are worthy of study. Contemporary statistical thinking may not be fully compatible with the way scientists attend to outliers, which are often the stimulus to important developments in theory and are de-emphasized only if they are determined to be unimportant artifacts.

The Hertzsprung-Russell diagram remains useful in work that is at the forefront of astronomical research and continues to inspire new questions regarding the origin and structure of the universe.

[*Received June 1990. Revised February 1992.*]

REFERENCES

Baade, W. (1944), "The Resolution of Messier 32, NGC 205, and the Central Region of the Andromeda Nebula," *Astrophysical Journal*, 100, 137–146.

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988), *The New S Language*, Pacific Grove, CA: Wadsworth and Brooks/Cole.

Beninger, J. R., and Robyn, D. L. (1978), "Quantitative Graphics in Statistics: A Brief History," *The American Statistician*, 32, 1–11.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

Cleveland, W. S., and McGill, R. (1984), "The Many Faces of a Scatterplot," *Journal of the American Statistical Association*, 79, 807–822.

Davis Philip, A. G., and Hayes, D. (1978), *The HR Diagram*, Dordrecht, Holland: Reidel.

Draper, N. R., and Smith, H. (1981), *Applied Regression Analysis* (2nd ed.), New York: John Wiley.

Friedman, J. H., and Rafsky, L. C. (1981), "Graphics for the Multivariate Two-Sample Problem," *Journal of the American Statistical Association*, 76, 277–287.

Garrison, R. F. (1988), "Comments on the Method of Spectroscopic Parallax," *Publications of the Astronomical Society of the Pacific*, 100, 1036–1040.

——— (1992), "Stars, Hertzsprung-Russell Diagram," in *The Astronomy and Astrophysics Encyclopedia*, New York: Van Nostrand Reinhold, pp. 736—739.

Hertzsprung, E. (1905), "Zur Strahlung der Sterne," *Zeitschrift fuer Wissenschaftliche Photographie, Photophysik und Photochemie*, 3, 429–442.

——— (1907), "Zur Strahlung der Sterne II," *Zeitschrift fuer Wissenschaftliche Photographie, Photophysik und Photochemie*, 5, 86–107.

Houk, N., and Fesen, R. (1978), "HR Diagrams Derived From the Michigan Spectral Catalog," in *The HR diagram*, eds. A. G. Davis Philip and D. S. Hayes, Dordrecht, Holland: Reidel.

Lambert, J. H. (1779), *Pyrometrie*, Berlin: Babey Houde und Spener.

Lewandowsky, S., and Spence, I. (1989), "Discriminating Strata in Scatterplots," *Journal of the American Statistical Association*, 84, 682–688.

Mosteller, F., Siegel, A. F., Trapido, E., and Youtz, C. (1981), "Eye Fitting Straight Lines," *The American Statistician*, 35, 150–152.

Russell, H. N. (1913), " 'Giant' and 'Dwarf' Stars," *The Observatory*, 36, 324–329.

Russell, H. N. (1914), "Relations Between the Spectra and Other Characteristics of the Stars," *Popular Astronomy*, 22, 275–294, 331–351.

Sandage, A. (1957), "Observational Approach to Evolution. II. A Computed Function for K0-M2 Stars From Mv = +5 to Mv = −4.5," *Astrophysical Journal*, 125, 422–444.

Smith, E. V. P., and Jacobs, K. C. (1973), *Introductory Astronomy and Astrophysics*, Philadelphia: W. B. Saunders.

Sneath, P. H. A., and Sokal, R. R. (1973), *Numerical Taxonomy*, San Francisco, CA: W. H. Freeman.

Spence, I., and Lewandowsky, S. (1990), "Graphical Perception," in *Modern Methods of Data Analysis*, eds. J. Fox and S. Long, Beverly Hills, CA: Sage, pp. 13–57.

Tufte, E. R. (1983), *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

Wainer, H., and Thissen, D. (1979), "On the Robustness of a Class of Naive Estimators," *Applied Psychological Measurement*, 4, 543–551.

Weisburg, S. (1980), *Applied Linear Regression*, New York: John Wiley.