Using scatter diagrams to study cause-and-effect relationships

Alfred H. Jaehn

The simplest way to begin a study of cause-and-effect relationships is to prepare a scatter diagram. It is always easier to see the relationship in a scatter diagram than in a table of numbers. Paired data of the variables to be examined are obtained and plotted on graph paper. The vertical (y) axis is reserved for the characteristic to be predicted (e.g., Scott Bond, opacity, tensile strength). The horizontal (x) axis is used for the variable that is being used to make the prediction (e.g., percent wet-end starch, parts titanium dioxide in coating, chemical/mechanical pulp ratio).

The relationships shown on a scatter diagram can be studied nonmathematically by observing the pattern of the plotted points. The correlation is said to be positive when the y value increases as the x value increases. If the y value decreases as the x value increases, the correlation is said to be negative. A random pattern indicates a lack of correlation.

Figures 1–3 illustrate typical scatter diagrams. Figure 1 shows the presence of a positive correlation. On the other hand, Fig. 2 pictures a negative correlation. Finally, Fig. 3 fails to indicate any correlation at all.

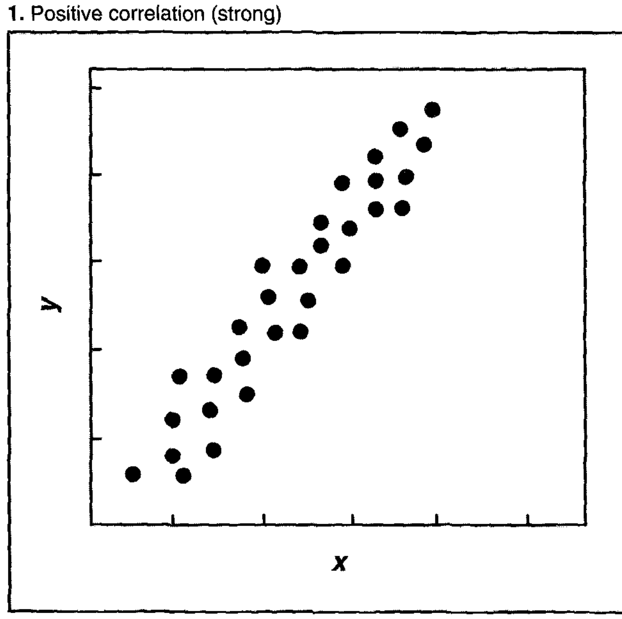
The degree of correlation is also readily evident on a scatter plot. Assuming the scales in Figs. 1 and 2 are identical, Fig. 1 shows a better correlation than Fig. 2. Note that the greater the scatter of points, the less the correlation.

Insufficient spread—a common pitfall

A frequent problem encountered in using scatter diagrams to correlate problems with potential causes is the limited range or spread of the data. For example, when attempting to show how a product characteristic was affected by a process or formulation change, a conclusive correlation is often not obtained. Although our technical knowledge supported by previous laboratory experiments tells us that a relationship should exist, this is not evident when plotting the mill data. In this case, the range of the production conditions was far less than that employed in the laboratory. Only by widening the range of the mill operating conditions can the true cause-and-effect relationship be seen.

Jaehn is a private consultant in quality management for the process industries, 1430 Clyde Ave., Wisconsin Rapids, WI 54494. He is an ASQC Certified Quality Engineer, TAPPI Fellow, and an Examiner for the Malcolm Baldrige National Quality Award.

1. Positive correlation (strong)

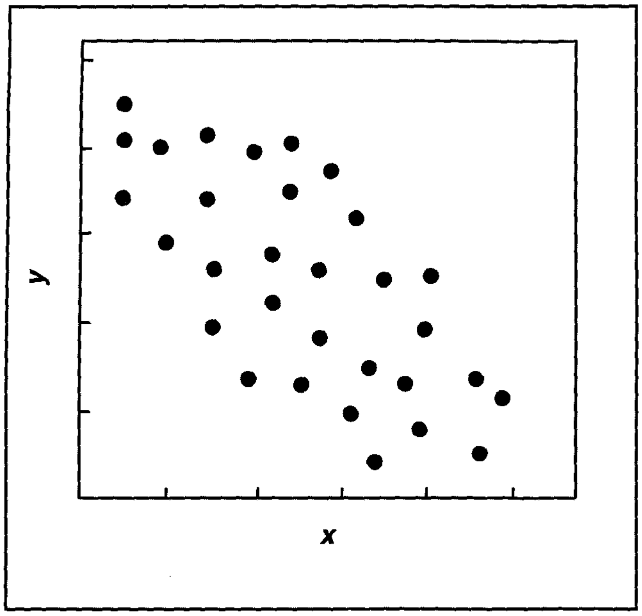


Difficulties in establishing relationships also often occur when attempting to correlate mill test data. In many cases, it is advisable to extend the range of the data by incorporating results from other sources (including a wider range of products, manufacturing conditions, etc.). Increasing the spread of the data helps compensate for differences associated with the exact matching of test data with mill operating conditions, sampling problems, and measurement errors. **Figure 4** illustrates this point. This plot describing the effect of basis weight on tensile strength shows that if the range were limited to one specific grade and weight, the resulting data would indicate that no correlation exists. However, by including the same grades produced at different basis weight targets, the existence of a correlation becomes obvious as shown in **Fig. 5**.

Improving correlations by separating the data

Identifying factors associated with cause-and-effect relationships is frequently helpful in problem-solving activity. Separating the data by specific area, equipment, operator, raw material being used, etc. can often provide insight as to why an expected correlation was either not obtained or very poor.

2. Negative correlation (weak)



3. Lack of correlation

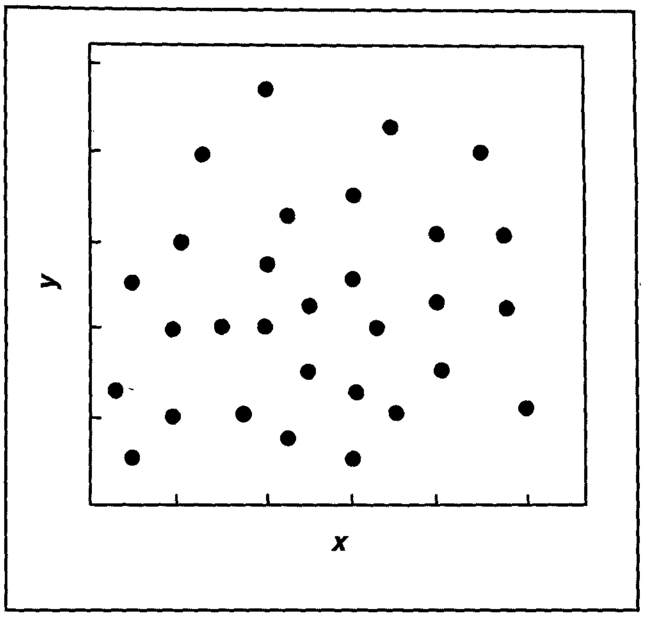
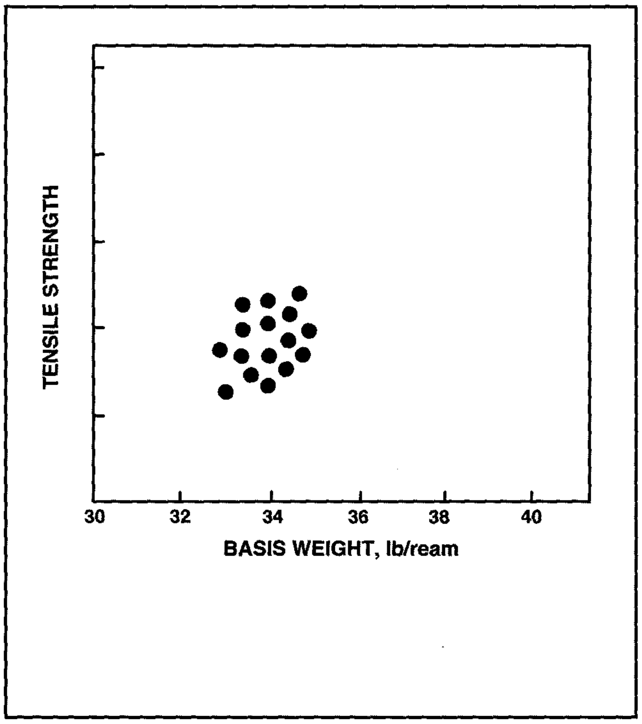


Figure 6 provides an example of the advantages of separating the data. These data are the identical tensile strength results vs. basis weight results plotted in Fig. 5. In this case, the results were separated by the pulp source in use at the time the paper being tested was manufactured. Clearly, the correlation between tensile strength and basis weight is improved (reduced scatter) by separating the data by pulp source.

The process of separating the data is referred to as stratification. The selection of the stratification variables is often constrained by the available data. After all, one cannot expect that the desired facts relevant to every set of data would be included in the mill record keeping system. When there is insufficient information on one or more of the identifying variables, it is necessary to plan a program for its collection. If new data are to be obtained, care must be exercised to assure that all useful identifiers are included.

4. Insufficient data spread



Some useful tips

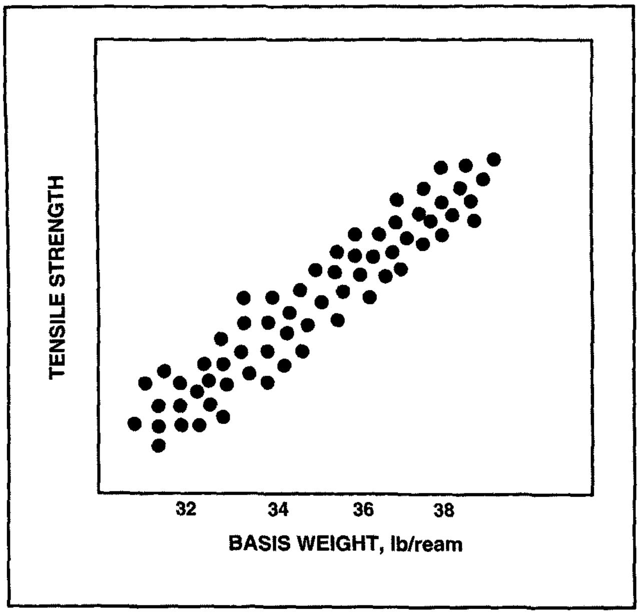
Scatter diagrams only picture the relationships between variables. They do not necessarily prove cause and effect. Sometimes it is possible that a correlation seen on a scatter diagram is due to a cause other than the one being plotted. One must still use fundamental knowledge and understanding of the process being examined when drawing conclusions from scatter diagram appearance.

When preparing a scatter diagram, attention needs to be given to the scaling of the x and y axis. It is generally advisable that the maximum and minimum values of each axis closely correspond to the maximum and minimum value of each variable. If this is not consistently done, the opportunity of finding a correlation may be missed entirely. If the same predicted characteristic is to be plotted vs. several possible causes, then the y axis should be scaled the same to avoid any bias in the final comparisons.

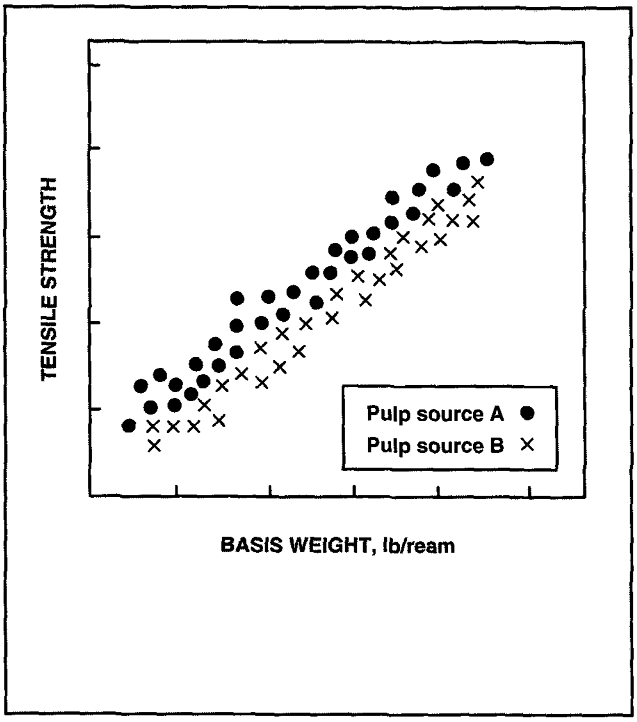
Attention also needs to be given to the number of paired data points used in plotting a scatter diagram. About 30 pieces of data are generally sufficient. However, if a correlation is very strong, 10–20 points may be adequate. Also, when data are separated into categories, one needs to assure that there are enough data points in each group to provide a meaningful analysis. Again, at least 10–20 points in each identified category should be considered as a minimum requirement.

Finally, one needs to be cautioned about "eyeballing" a

5. Spreading out the data

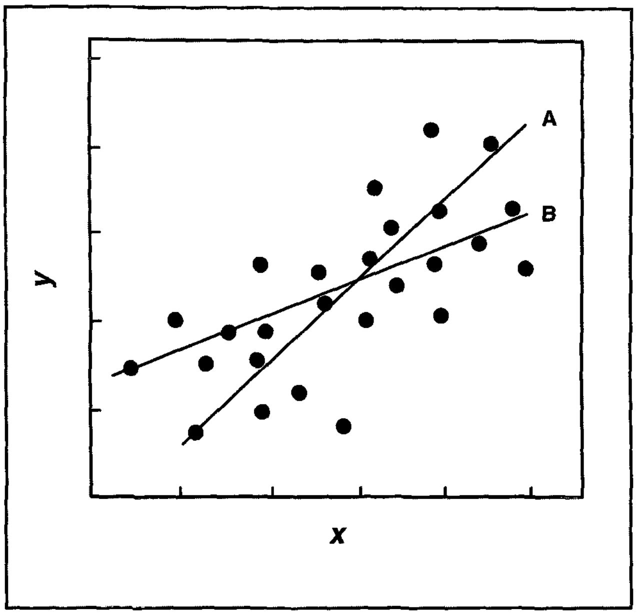


6. Separating the data



best fit line to represent the correlation between two variables. Unless the line is determined mathematically by regression analysis, it should not be used to draw conclusions about the cause-and-effect relationship. This is particularly important if a large degree of scatter is present. **Figure 7** illustrates two supposed "lines of best fit" on a scatter diagram. Which of these would you choose to represent the relationship?

7. Choosing the "best fit line"



Closing remarks

Scatter diagrams provide a "picture" of the data. The general relationship between two variables can be examined immediately—to determine if it is strong, weak, or nonexistent. They do not, however, prove cause and effect. Scatter diagrams are only a starting point in the problem-solving journey. Only after reviewing the knowledge at hand and performing designed trials can one boldly proclaim that a true cause-and-effect relationship exists. □

Received for review April 3, 1991.
Accepted April 18, 1991.