CC/NUMBER 11
MARCH 17, 1986

This Week's Citation Classic®

Lance G N & Williams W T. Mixed-data classificatory programs. I. Agglomerative systems. *Austral. Comput. J.* 1:15-20, 1967.
[CSIRO Division of Computing Research, Canberra, ACT, Australia]

This was the first in a series of papers on mixed-data classificatory programs. It dealt with agglomerative hierarchical methods and showed how an information statistic can be used with mixed data. [The *SCI*® indicates that this paper has been cited in over 145 publications, making it this journal's most-cited paper.]

W.T. Williams
10 Surrey Street
Townsville 4812
Australia
and
G.N. Lance
Department of Computer Science
University of Bristol
Bristol BS8 1TW
England

October 23, 1985

The year 1966 was something of a milestone in the progress of numerical classification. Methods had existed in theory for many years, but they were out of reach of hand computation and far too time-consuming even for first-generation computers. The advent of second-generation computers promised that it might at last be possible to solve real-life problems in classification. As a result, all over the world—especially in Australia, England, and America—people began to write programs for numerical classification. However, anybody wishing to enter this field, or even simply to use the programs, was faced with two difficulties. The first was that virtually all existing programs could only handle a single type of data—usually all-binary or all-numeric. Yet real-life problems were not like this. Biologists, for example, often recorded a collection of several different types of attribute. All that could then be done was to force their data into one of the types acceptable to one of the programs. We had already pointed out, "The day should be over when it is necessary to distort a taxonomist's original data for the convenience of the computer," and we were determined to devise algorithms that could process virtually any type of attribute. The second difficulty was that of choice between competing programs, all of which had been devised independently of each other. It occurred to us that there ought to be a general theory of such programs, of which all existing variants were simply special cases.

In the paper under discussion we gave the first account of a general theory of those systems known as "agglomerative hierarchical classifications," the commonest type. Papers on other types were to follow later.[1,2] The paper also provided algorithms for a wide variety of attributes and gave an account of the first mixed-data program using an information statistic.

The paper was well received at the time; controversy was to follow a year or two later.[3,4] Statisticians had never liked numerical classification because it did not test a hypothesis and so did not emerge with a probability. Mathematicians considered our algorithms crude—which they were, but they did the job required of them. The mathematicians later conceded that perhaps people really did want to classify things, so they produced an algorithm that nobody used (in our view because they had solved, with great elegance, the wrong problem).

The paper is, of course, now out of date in several respects and has been superseded by more complete and erudite reviews. So why is it still cited? We think there are two reasons. First, it is accessible; later reviews have tended to be in user journals that computing departments may not carry. Second, a form of historical courtesy appears to be operating—a feeling that, if one is going to cite *any* general account, it would be polite to cite the first.

1. **Lance G N & Williams W T.** Mixed-data classificatory programs. II. Divisive systems. *Austral. Comput. J.* 1:82-5, 1968.
2. **Lance G N, Milne P W & Williams W T.** Mixed-data classificatory programs. III. Diagnostic systems. *Austral. Comput. J.* 1:178-81, 1968.
3. **Sibson R.** Some observations on a paper by Lance and Williams. *Computer J.* 14:156-7, 1971.
4. **Williams W T, Lance G N, Dale M B & Clifford H T.** Controversy concerning the criteria for taxonometric strategies. *Computer J.* 14:162-5, 1971.

DA-8