Geographical Systems, Vol. 4(3), pp. 277–280
Reprints available directly from the publisher
Photocopy permitted by license only

ALGEBRA AND INFORMATION LOSS: A RESPONSE TO KUHN

MICHAEL F. GOODCHILD*

National Center for Geographic Information and Analysis and Department of Geography, University of California, Santa Barbara, CA 93106

Werner Kuhn's paper "Approaching the Issue of Information Loss in Geographic Data Transfers" makes a fascinating connection between the abstract algebra of groups and the apparently mundane problem of transferring data between geographic information systems (GIS). I have two related purposes in the following comments on his paper: first, to try to place a provocative and stimulating paper more firmly within the broader context of cartography and GIS; and second, to present an alternative perspective to what is undoubtedly a complex and difficult issue.

What do we mean by "information loss"? In recent years, cartographers have begun to emphasize the role of use in the definition of data quality: Moellering, for example, in a background paper to the U.S. Spatial Data Transfer Standard, now adopted as Federal Information Processing Standard 173, writes that "The purpose of a Quality Report is to provide detailed information for a user to evaluate the fitness of the data for a particular use" (Moellering, 1987, p. 8; see also Guptill and Morrison, 1995). More broadly, information is lost if a user is no longer able to perform a given operation, or manipulate the data in some way to satisfy a given purpose.

This comparatively recent trend in the cartographic literature echoes a much more highly structured trend that has occurred in the theory of computing, and that underlies Kuhn's paper. Data by itself is no more than an apparently random collection of bits—its meaning is defined by the operations that can be performed on it. If operations are no longer possible, or

* Corresponding author. Tel.: +1 805 893 8049. Fax: +1 805 893 7095. E-mail: good @ncgia.ucsb.edu.

give different results, then information has been lost. This provides a rigorously defined, well-structured, and potentially powerful way of characterizing the effects of a transfer of data from one system to another.

But while the approach may be rigorous, and therefore amenable to mathematical analysis, it is not necessarily simple, readily implemented, or readily understood. Nor is it unique—statistics, for example, provides a framework for the assessment of information loss which is also frequently compelling. We commonly measure the loss of information that is due to such processes as filtering, addition of noise, or measurement error using simple statistical indices that may be embedded in turn within rigorous statistical theory.

The power of an approach based on operations lies in part in its generality. It proposes, for example, that throwing away bits or digits in a transfer only amounts to information loss if their content cannot be determined in some other way from the remaining data. But while the approach may be rigorous and general, some of its consequences may be surprising, if not counter-intuitive. For example, consider Kuhn's Figure 4. Systems B and C both store names and point coordinates, and their databases are therefore identical in content. But information has nevertheless been lost in the transfer from B to C, according to this definition of loss, because C is unable to perform one of B's functions (compare points by name). If we transfer data from B to C, and then back to B, we are left with the paradox that two successive transformations, both involving loss of information, can replicate the original perfectly. This makes perfect sense within the somewhat abstract structure of the analysis, but its intuitive sense seems less than perfect. Intuitively, a transformation from System B to System A, which involves discarding names and is therefore irreversible, seems a different and more severe kind of information loss than one from System B to System C. Discarding non-redundant bits seems more drastic than placing the data in a system which lacks a particular function, particularly given the rapidly changing nature of most practical GIS environments, and trends in the software industry towards greater flexibility and modularization.

This argument may appear somewhat contrived, if it is interpreted as pitting a rigorous approach, well-grounded in algebraic theory, against one that is based on intuition. Thus the remainder of these comments review the arguments within the context of Kuhn's theoretical framework.

In group theory, a group consists of a non-empty set and an operation on members of the set that together satisfy certain properties (the definitions in this section are drawn from Doerr and Levasseur, 1985). The result of applying the operation to members of the group is called its product. An example of a group is the set of positive real numbers and the operation of multiplying pairs. Call this group G. Suppose one had a system (mental arithmetic) that was good at adding, but multiplication was generally impractical. In other words, the system represented by group G is not practical, but an alternative system G′, consisting of the set of real numbers and the operation of adding pairs, is both practical and readily available. Then the problem of practical multiplication could be solved if some transformation F could be found to take a pair of elements of G, convert them to elements of G′, perform addition, and make another inverse transformation back to G. Of course, the logarithm function is exactly that transformation, and its inverse is readily available.

The log transformation is an example of the property of isomorphism, and groups G and G′ are said to be isomorphic. An isomorphism must be 1:1, so there is a unique element in G′ for each element in G. It follows that it can also be inverted. In practical terms, an isomorphic transfer from System A to System B will be lossless because the transfer can be inverted to recover exactly the original information.

Kuhn uses the logarithm function as an example of the less restrictive property of homomorphism. In a homomorphism it is not necessary for the outcome of the transformation to be unique, and it is therefore not generally possible to invert the transformation F. A function F is said to be homomorphic if the result of applying F to the product of G is the same as if F had been used to transform elements of G to G′, and then the product of G′ had been computed. But it is perfectly possible for many different elements of G to transform to the same element of G′.

This is a fundamental difference, and it leads to a significant practical problem. Consider Kuhn's Figure 1. In the notation of these comments, the group G is the combination of Figure 1's Domain A1 and function f; G′ is the combination of Domain A2 and function g; and the transformation F is both mappings h1 and h2, which must be equal by definition (the notation has been changed in these comments to facilitate references to group theory). Practically, we would want to show that by using System B we can obtain the same results as with System A. But in a homomorphism the inverse of transformation F does not normally exist. In the logarithm example, this would be the equivalent of not having an antilog function—we would be left without the product of two numbers that we originally wanted.

The less restrictive nature of homomorphism leads to another significant practical implication (Doerr and Levasseur, 1985, p. 414). It is often important to know that some function F exists that satisfies the isomorphic property for two groups G and G′, because that suggests that G′ can usefully substitute for G, as in the logarithm example above, and that information can be transferred from G to G′ and back to G without alteration. But many homomorphisms may exist between two groups, many of them trivial (in the case of G and G′ defined above, the function "set to zero" is a homomorphism, since the result of zeroing a product is always equal to the sum of two zeroed values). The homomorphisms between two structures define the degree to which one structure is a model of the other, and the properties that it retains.

Both properties, of isomorphism and homomorphism, are defined at a level of reduction that may seem absurd to many spatial analysts and GIS users, since they define information loss at the level of the individual operation rather than the entire system. Take the equivalent of the multiplication problem in the context of GIS—suppose I have data in System A, but lack a certain function. According to this framework, I would search over all available Systems B, testing each function of each system to see if it, in combination with a forward and inverse transformation F, satisfied the isomorphism property. By comparison, the degree of invertibility of a transformation, based on a simple comparison of a data set with the results of transferring to some other system and then back again, seems a simple, intuitive, readily computed, and comprehensive measure of the information lost in using that system as a substitute. On the other hand, it is not specific to operations, and thus fails to deal with the basic problem addressed in Kuhn's paper—the need to include function, and therefore use, in the definition of loss.

Outside the relatively narrow confines of group theory, "isomorphic" and "homomorphic" convey a sense of "equal" and "similar" respectively. These comments have raised questions about the appropriateness of both terms to the problem of information loss in geographic data. Whether they turn out to be (broad sense) isomorphic or (broad sense) homomorphic, and whether something of practical value can be extracted from the relationship, remains to be seen. But whatever the outcome it is clear that Kuhn's paper has raised interesting and challenging issues.

References

Doerr, A. and Levasseur, K. (1985) *Applied Discrete Structures for Computer Science*. Science Research Associates, Chicago.

Guptill, S. C. and Morrison, J. L., editors (1995) *Elements of Spatial Data Quality*. Elsevier, New York.

Moellering, H, editor (1987) A Draft Proposed Standard for Digital Cartographic Data. Report No. 8, National Committee for Digital Cartographic Standards, American Congress on Surveying and Mapping, Bethesda, MD.