# THE NATURE AND VALUE OF GEOGRAPHIC INFORMATION

Michael F. Goodchild[1]

ABSTRACT

If geographic information is to be traded as a commodity, it must be possible to measure its quantity, to compare the contents of two data sets, and to perform other simple functions such as the measurement of similarity of content. Shannon-Weaver information theory is found to be inadequate for this purpose. In current practice geographic information is priced by such medium-dependent measures as area, or data volume. A theoretical framework is proposed in which geographic information is defined by atomic tuples, to which other structures can be reduced. Although the potential number of such tuples is infinite, the ubiquity of spatial dependence and the impossibility of determining location exactly allow useful representations to be created from finite numbers of tuples. Based on this definition of geographic information, formal definitions are proposed for geographic information systems and geographic queries. It proves difficult to incorporate naïve geography within this framework because of its logical inconsistencies. A semantic theory of geographic information is proposed, based on networks of linkages expressed as tuples. Unresolved questions are identified.

## 1. INTRODUCTION

Markets for commodities such as wheat, crude oil, or pork bellies depend on the ability to measure quantity, whether it be in bushels, barrels, or tons. On the other hand information-rich commodities such as books or newspapers are priced per unit rather than by such measures as volume (despite apocryphal stories of assessing dissertations by weight), since they are always bought and sold whole, and there is little point in trying to establish a price for half a book, or a fraction of a newspaper. Similarly paper maps are priced by unit, rather than by square kilometer or some other continuous-scaled measure.

One of the underlying principles of the information economy is that digital information can also be traded as a commodity. But for some types of information the transition to digital form undermines the concept of a unit, and with it the information's basic granularity. Geographic information is a case in point, since digital geographic information is often merged into seamless databases, and disseminated for user-specified areas. A user might be interested in purchasing only the data covering a circular area surrounding a point, or only the data covering an irregularly shaped county or school district. In such cases a market for geographic information requires well-defined measures of information quantity, because otherwise all transactions would be unique, and no market could exist. For example, it must be possible to establish the combined

[1]National Center for Geographic Information and Analysis, and Department of Geography, University of California, Santa Barbara, CA 93106-4060, USA. Phone +1 805 893 8049, FAX +1 805 893 3146, Email good@geog.ucsb.edu

value of two data sets given the values of each independently, using a rule such as a simple summation that obeys formal mathematical principles.

Buyers and sellers of digital geographic information have struggled with this issue, but little progress appears to have been made. The research initiative of the U.S. National Center for Geographic Information and Analysis on the Use and Value of Geographic Information discussed this many related issues (Onsrud and Calkins, 1993) without reaching definite conclusions, and other more recent reviews have similarly failed to report substantive progress (*e.g.*, Longley *et al.*, 2001). Instead, the community has adopted such measures as data volume (in bytes), and geographic area of coverage. But it makes little sense to price information by volume if such measures are sensitive to simple manipulation. For example, suppose a 15MB data set is compressed to 3MB using a loss-less compression such as run-length encoding; is it now worth only 20% of its previous value? What if a vector data set is restructured in raster form, and as a result grows in volume by a factor of ten? Moreover, pricing by area of coverage makes little sense if the volume of information, and the cost of collecting it, varies dramatically from one area to another.

A satisfactory measure of information quantity would have to allow price to be determined for any user-defined area. It would have to be independent of data structure, since simple restructuring (*e.g.*, from one data format standard to another) should not affect price. It would have to be independent of medium, if price is to be determined by data content rather than by the medium on which the content is stored. Methods should also be available to determine if two data sets have the same content, and differ only by such comparatively irrelevant factors as medium, or data structure, or projection. One of the advantages of digital information is the ease with which it can be mutated into other forms, through simple transformations. A transformation is said to be *reversible* if its inputs can be recovered by a simple reverse transformation. Today's geographic information systems make it easy to perform a range of reversible transformations, including changes of projection and datum, or topological overlay. A satisfactory measure of information quantity clearly should not be affected by reversible transformation (see Kuhn, 1997, for a comprehensive discussion of this issue).

These requirements can only be achieved through a theory of geographic information that is *semantic* rather than *syntactic*; in other words, one that focuses on the meaning of information rather than on its form. A semantic measure of information should also distinguish between information that adds to the receiver's knowledge, and information that merely duplicates existing knowledge; in the second case the information content should be zero. A semantic measure of information must therefore depend not only on the content of the message, but also on various aspects of the state of the receiver's knowledge, including knowledge of relevant rules and conventions.

This chapter is intended to make a few small steps in the direction of such a semantic measure, and associated theory. It begins with a discussion of classical syntactic information theory, in order to demonstrate its inadequacy. A theory is proposed in which geographic information is constructed from simple atomic tuples, and the problems of

dealing with the potentially infinite number of such tuples are addressed. The subsequent section addresses queries, and the role of geographic information systems (GIS), within this theoretical framework. This is followed by a discussion of two issues of direct relevance to the framework: digital Earth, and naïve geography. The framework is used to develop a semantic theory of communication that satisfies the requirements outlined above. Finally, the chapter concludes with a short discussion of unaddressed issues.

Many aspects of the approach are general, and might be usefully applied to any type of information. But geographic information seems a particularly well-defined type or subset of information, and as such provides an excellent base for speculating about the nature of information in general.

## 1.1 SHANNON-WEAVER INFORMATION THEORY

Syntactic theories of information have been developed extensively over the past half century, largely in the context of coded communication. Shannon (1949) proposed that the information content of a coded message could be measured in terms of the probabilities of occurrence of each possible code. Suppose, for example, that a message is sent in binary code as a string of 0s and 1s, and that in such messages the digits 0 and 1 occur with roughly equal frequency. Then a single digit in the message can be regarded as resolving between two equally likely possibilities, and its information content can be measured accordingly using metrics of the form $-p_i \log p_i$, where $p_i$ is the probability associated with code *i*. Logs to base 2 are convenient when the code is binary, and in this case the units of the measure are known as *bits*. The measure is designed to remain constant despite a recoding of the message into some other coding scheme.

It is easy to extend this approach to more complex codes. For example, consider a single letter of the Roman alphabet and a message in the English language. Each letter now resolves among 26 possibilities, but in plain English the probabilities associated with each letter range widely, since E occurs much more frequently than X, and consequently can be said to convey less information. Further complications arise when sequences of more than one letter are considered, since the information content of a U following a Q in English is almost nil (its conditional probability is almost 1).

Certain forms of geographic information might be amenable to such an approach to measuring information content. Suppose a map of land cover type is coded using a raster structure, and each cover type *i* is associated with a probability $p_i$. Then the information content of the entire raster might be measured by treating it as a linear series of codes in the alphabet defined by the set of cover types. But it would be difficult to capture the effects of conditional probabilities, since in a typical geographic context a pixel surrounded by pixels of type A is more likely to be type A than any other type, and thus to have less information content. The approach can be applied hierarchically, to determine the information content of maps at different scales (for reviews of the use of information statistics in geographic research see Marchand, 1972; Thomas, 1981). In this context the use of information statistics is directly comparable to decompositions based

on variance (*e.g.*, Csillag and Kabos, 1996) or to Fourier or wavelet decompositions. One might similarly attempt to apply this approach to printed maps, based on the probability that any one point on the map is covered by any one of the map's inks. But here the number of points, and hence the length of the message, is potentially infinite. Moreover, it would be wrong to assess the points comprising a letter of annotation (such as the black ink denoting the letter A) as a set of independent codes, since they can also be regarded as a single code resolving among the 26 letters of the Roman alphabet.

This last problem is well illustrated using another simple example. Consider an infinite series of digits beginning with 3.14159… As an infinite series of apparently random decimal digits it has infinite information content. But the same information might be communicated using a single letter π from the Greek alphabet. What appears to be an example of infinite data compression is of course enabled by the convention that uses that particular Greek letter to denote that particular infinite series, and if the convention is not known to the receiver of the message then the point is lost and the message has no value.

Shannon-Weaver information theory is concerned with the form and coding of messages, and clearly fails to satisfy the requirements laid out above for a semantic theory of information. Such a theory would deal effectively with the last example, and recognize the equivalence of, first, an infinite transmission of digits; and second, the transmission of a Greek letter coupled with the knowledge that the receiver understands the relevant convention. It might also recognize a third case: the transmission of a computer code or mathematical rule for determining the digits of π. In principle, all of these three possibilities should have the same information content in a semantic theory of information.

## 2. A THEORY OF GEOGRAPHIC INFORMATION

Several attempts have been made to develop general theories of geographic information. Berry (1964) described a *geographic matrix*, a general model in which geography is characterized by location, time, and attributes. The same idea underlies Sinton's conceptualization (Sinton, 1978), as well as numerous discussions of geographic data modeling (*e.g.*, Peuquet, 1994).

Goodchild *et al.* (1999) propose that geographic information can be defined by reference to an *atomic element* or tuple of the form <**x**,**z**> where **x** denotes some location in space-time, and **z** denotes some set of properties associated with that location, commonly termed attributes. Location is constrained to the Earth's surface and near-surface. In this framework, all geographic information is ultimately reducible to a set of such atoms. For example, a digital elevation model consists of a series of atoms denoting the elevation at a regularly spaced series of points; elevation is the property associated with those locations; and time is ignored because elevation is assumed to be a static property, tectonic and other processes notwithstanding.

No message would be of any value to a receiver who did not understand the conventions underlying the message. In the case of geographic information, these include the conventions associated with position on the Earth's surface, such as latitude and longitude, datums and projections; and those associated with time, and we term these *universal locators*. With such universally accepted standards it is reasonable to expect that a receiver would understand **x** in most cases. But there is far less universality associated with **z**. While the Celsius scale of temperature is universal in the scientific world, the general public are more likely to use such vague terms as *warm* or *cold* to describe the temperature of the atmosphere near the ground. Other commonly encountered attributes such as land cover class are far less standardized, and the potential exists for far more confusion over the meaning of messages.

The dimensions of space and time that define **x** are continuous, and an infinite number of locations therefore exist. Thus while information gathered at points can be represented by individual atoms, an infinite number of atoms is required to characterize variation over lines, areas, or volumes. For example, it would require an infinite number of atoms to characterize the spatial variation of elevation over a finite area, or to define the extent of the State of California.

Two principles work to address this issue, and to make it possible to characterize geographic variation in a finite number of atoms. The first is often known as Tobler's Law of Geography (1970): "All things are related, but nearby things are more related than distant things." The effect is often measured using indices of spatial dependence, such as the Moran or Geary indices of spatial autocorrelation (Cliff and Ord, 1981) or the variogram (Isaaks and Srivastava, 1989), that compare variation over different distances. For example, the variogram plots average or expected variance between pairs of observations against the distance between them, and is normally observed to be a monotonically increasing function. It is easy to see the validity and generality of Tobler's Law if one tries to imagine a world in which it is not true. Such a world would be impossible to describe or inhabit, since the full range of variation could be encountered over vanishingly small distances.

In practice, many geographic phenomena exhibit constant values, or variation that lies below some acceptable threshold, over finite neighborhoods. The property "State of California" is uniformly true of all of the points within the state's boundaries, and similar principles apply to the attributes relevant to land ownership. In the case of land cover, class is held to be approximately constant within areas mapped as uniform class. For properties measured on continuous scales, such as elevation, Tobler's Law implies relationships of the form $\left|z(\mathbf{x}+\delta\mathbf{x})-z(\mathbf{x})\right|<\lambda$ for $\delta\mathbf{x}<\tau$ where $z$ is such a property, $\tau$ defines a neighborhood in space and time, and $\lambda$ defines an acceptable threshold of difference.

In other cases, Tobler's Law is implemented in the form of a series of *interpolation rules* that allow atomic tuples to be inferred from surrounding ones. For example, if temperature is known at a series of weather stations, then temperatures are inferred at

intervening locations through simple processes such as Kriging or inverse-distance weighting (Longley *et al.*, 2001).

The second principle derives from the process of determining location **x**. Location on the Earth's surface can be determined by a variety of means, including surveying, the use of the Global Positioning System, or by reference to a paper map. In each case the result is subject to measurement error, because of the limitations of the measuring instruments. Moreover, difficulties in defining the exact shape of the Earth, the tendency of the reference poles to wobble, and persistent crustal movements all work to ensure that measurement errors will never be reduced to zero. In practice, therefore, it is sufficient to represent geographic variation over a finite rather than an infinite number of points.

Moreover, the spatial resolution required by any application, or obtainable from any data acquisition system, is also strictly limited. Suppose a spatial resolution of $\lambda$ is sufficient for a given application, $\lambda$ being measured in linear units. Assuming a spherical Earth of radius *R*, the number of atoms required to characterize the spatial variation in some phenomenon over the surface of the Earth is approximately $4\pi R^2 / \lambda^2$ . The same number would be appropriate if $\lambda$ instead represented the inherent spatial accuracy of the measurement of position (see, for example, Güting and Schneider, 1995).

The number of atoms needed to characterize the geographic domain is also dependent on the complexity of **z**, and the number of distinct properties present at any location. In principle an infinite number of properties might be held to exist, but in practice the number is finite, and strong correlations exist between many properties.

## 2.1 FIELDS AND DISCRETE OBJECTS

Because of these principles, and their effect in limiting the number of atoms, it has been possible to construct satisfactory representations of many geographic phenomena. Two approaches are used, representing alternative conceptualizations of variation in space (Worboys, 1995). The first, or *field* perspective, conceives of geographic variation in terms of a number of variables that are functions of position, $z = f(\mathbf{x})$. This conceptualization fits many physical phenomena, including land surface elevation, and atmospheric temperature and pressure. From the previous section, we know that a finite number of atoms is in practice sufficient to characterize a field with adequate accuracy over a finite space-time domain, and six approaches are commonly used in the case of two spatial dimensions and notemporal variation, as shown in Figure 1.

[Figure 1 about here]

Row by row from the top left, the six approaches are as follows:

1. A finite number of regularly spaced points is used to create a finite number of atoms. Values at other locations are determined by an interpolation rule (for example, the rule that *z* is the value at the closest data point).

2. A finite number of irregularly spaced points is used. Values at intervening points are determined by an interpolation rule such as Kriging or inverse-distance weighting.

3. The area is partitioned into a finite number of rectangular regions, and a single value is recorded for each region. Detailed variation within regions is ignored.

4. The area is partitioned into a finite number of irregular polygonal regions, each defined by a number of points, and a single value is recorded for each region. Detailed variation within regions is ignored.

5. The area is partitioned into a finite number of irregular triangular regions. Atoms are recorded at the vertices, and values are interpolated within regions using linear functions (this approach can be generalized to include quadrilateral regions and nonlinear functions).

6. A finite number of isolines is defined. Each isoline is represented by an ordered set of points connected by straight segments, and a single value is recorded for each isoline (this approach can be generalized to allow for curved segments). Values between isolines must be obtained using an interpolation rule.

Each of these methods succeeds in characterizing the field, using a finite number of tuples. The form of the tuples varies; in (1), (2), and (5) the tuples are of the form $<\mathbf{x},z>$, while in (4) and (6) they are of the form $<x_1,y_1,x_2,y_2,x_3,y_3,\ldots,x_n,y_n,z>$, and in (3) a single tuple is sufficient, of the form $<z_1,z_2,z_3,\ldots,z_n>$, given a known rule for ordering the elements of the raster, and a basis for its referencing to the Earth's surface.

In the *discrete object* perspective, the geographic world is conceived as empty, except where it is occupied by one or more points, lines, areas, or volumes. Each of these discrete objects has attributes describing its properties. Points are described by single atoms, and lines and areas by tuples of the form defined for cases (4) and (6) above in the static, two-dimensional case.

These methods succeed in creating data structures that use a finite number of tuples to describe geographic phenomena. Each can be decomposed into atomic tuples; for example, an infinite number of distinct tuples can be derived from any of the field representations and associated interpolation rules.

## 3. GEOGRAPHIC QUERIES

The previous section described a theory of geographic information based on distinct atoms or tuples. This section builds on that base by considering the nature of geographic queries, and the systems that respond to them. The term *geographic information system* (GIS) is used here to refer to such systems, and it is important to recognize that while most references to GIS imply a digital system, this same approach might be applied to the processes humans employ when responding to queries.

A *simple geographic query* is defined as a query to which an atomic tuple of geographic information is the answer. Two types of query exist, of the forms "What is at **x**?" and "Where is **z**?", and both are answered by providing the remaining part of the appropriate atomic tuple (or tuples). Two more advanced forms of query can be defined: those that can be answered from more complex structures without access to the atomic form, such as "What are the properties of this area?"; and those that require deduction from the contents of two or more atoms, such as "What is the distance from A to B?", which can be deduced from $<\mathbf{x}_1,A>$ and $<\mathbf{x}_2,B>$.

It is now possible to define what is meant by the possession of geographic information, in a manner that is independent of medium and structure. A GIS is said to possess an item of geographic information (defined as one or more atoms) if *it is capable of responding to a query to which the item is the answer*. This definition allows for the possibility that the GIS must undertake some form of transformation, processing, or deduction in order to obtain the answer. Note also that the definition defines possession in the context of a GIS rather than independently, implying that the system and the information are effectively inseparable.

Given such a definition, one might wonder whether it is possible to measure the quantity of information by counting the number of distinct queries that can be answered, and thus in effect counting atomic tuples. Consider, for example, the query "Is **x** in California?" A GIS could address this query by performing a simple point-in-polygon operation on **x** to determine whether the point lies inside a polygon representing the boundary of the State of California. But clearly an infinite number of such queries exist, since an infinite number of distinct points can be defined on the Earth's surface, and there exists an answer to the query for each of them. Does the GIS therefore possess an infinite amount of geographic information?

A previous section has already addressed this dilemma in a different form. If position is not knowable to better than some length $\lambda$, or if the location of the boundary of California is similarly known only to a limited accuracy, then the number of distinct queries is finite. Moreover, if $\mathbf{x}_1$ is known to be in California and $\mathbf{x}_2$ is also known to be in California, then there is a high probability that $\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2$, $0 \le \alpha \le 1$ also lies in California (and a probability of 1 if California is convex), reducing the effective number of distinct queries further. Finally, the system is able to answer the query because it possesses a representation of the boundary of California as a polygon, plus an algorithm for resolving point-in-polygon queries. Thus the amount of information contained in the polygon definition could be said to place an upper bound on the amount of information possessed by the system, since the queries are resolved through a transformation of this information.

Another issue arises because of the possibility of answering queries corresponding to derivative information, rather than atomic tuples. For example, a system that *knows* $<\mathbf{x}_1,A>$ and $<\mathbf{x}_2,B>$ also is able to respond to a query about the distance between A and

B. But this issue is easily addressed by restricting queries to atomic forms, and thus measuring information by the number of atoms of specifically geographic information.

## 3.1 DIGITAL EARTH

The phrase *digital Earth* (DE) was coined by Gore (1992), and has been the focus of extensive discussion and the development of prototypes in recent years (see, for example, http://www.digitalearth.gov). DE is conceived as a repository of a vast array of information about the planet's surface and near-surface, in the future and past as well as the present, together with the means to render dynamic three-dimensional visualizations in virtual reality. In effect, it represents a single portal to all that is known about geography, that is, to all geographic information.

It is interesting to speculate on the nature of a complete DE. The previous discussion has shown that a finite number of atoms of geographic information can be sufficient to create a satisfactory representation, and Goodchild (2001) has argued that DE is technically feasible within the limits of today's Internet bandwidths and reasonable expectations concerning spatial resolution. Suppose, then, that one could build and maintain a DE, as a system capable of responding to any query about the surface and near-surface of Earth.

Now compare such a system to the one or more people able to observe and measure facts about the Earth's surface and near-surface through direct observation. Both DE and the observers would be equally capable of responding to geographic queries. Moreover, it would be impossible to design an experiment to determine whether a response to a query came from a GIS and DE, or from an observer – that is, to determine whether the responding system was in contact with *bit* or *it* (Siegfried, 2000).

## 3.2 NAÏVE GEOGRAPHY

The discussion thus far has been oriented to principles of scientific measurement, and has regarded the geographic world as a rigid, Newtonian frame. Science has always stressed the importance of shared understanding, and its role in communicating knowledge and in supporting the replicability of scientific results. But the human world does not necessarily respect such principles, and may instead be willing to accept the equal legitimacy of multiple, personal viewpoints. The research area known as Public Participation GIS (Craig, Harris, and Weiner, 2002) has investigated the ability of GIS to support multiple viewpoints; in the context of this chapter they might be regarded as multiple, equivalent versions of the same attribute, that is, $<\mathbf{x},z_1>$, $<\mathbf{x},z_2>$, $<\mathbf{x},z_3>$…

Egenhofer and Mark (1995) use the term *naïve geography* to describe a more human-oriented perspective, in which multiple points of view are possible, and in which the nature of geographic information reflects what people believe to be true, rather than what science has determined to be true. They define naïve geography as "the body of knowledge that people have about the surrounding geographic world."

Naïve geography creates interesting issues for the theoretical framework proposed here. On the one hand, multiple viewpoints can readily be accommodated through multiple tuples, as suggested above. But consider the statement "Santa Barbara is north of Los Angeles." The statement is widely believed to be true, and is reinforced by driving directions, since the main highway from Los Angeles to Santa Barbara is designated as U.S. Highway 101 North. But if one takes eight compass directions, or even four, in neither case is the bearing of Santa Barbara in the North sector (between 337.5 and 22.5, or between 315 and 45, respectively). Instead Santa Barbara is better described as west of Los Angeles.

One might imagine building a GIS that respects such beliefs, but in reality it would be impossible to do so, because the geometric rules on which GIS is based, and which permit such queries to be addressed, form a mathematical system that is consistent with a small number of axioms. A GIS that respected such beliefs could not reason, because information could no longer be reduced to atomic form, and the rules on which reasoning is based would no longer be general.

## TOWARDS A SEMANTIC THEORY OF GEOGRAPHIC INFORMATION

Consider the statement "Mount Everest is 8850m high". Symbolically, we might represent this as a tuple <"Mount Everest",8850>. But it fails to satisfy the earlier definition of an atomic tuple of geographic information because although 8850 is an instance of the attribute of elevation, "Mount Everest" does not directly identify a position on the Earth's surface. To satisfy the definition, it is necessary to know where Mount Everest is in some universal system of georeferencing, in other words to possess a second tuple <**x**,"Mount Everest">. Hence the tuple is of no value to someone who does not understand the term "Mount Everest", and is not geographic information until "Mount Everest" is converted into a universal locator. A *gazetteer* is defined as a collection of tuples linking placenames to universal locators.

The value of the tuple clearly depends on the receiver's knowledge of the location of Mount Everest, but it also depends on what else is known about Mount Everest, that location on the Earth's surface, and the elevation 8850. For example, one might know that the location is one of rapid crustal movement and uplift; that an elevation of 8850m places the summit in a zone dominated by strong jetstream winds; that summer precipitation in the area is linked to the Monsoon, etc. All of these facts can be symbolized as tuples linking either **x** or 8850 to other properties. Without them, the original tuple would be of no value; and the more of them there are, the greater the value of the tuple, and the more the tuple can support new reasoning, and successful response to new queries. The number of tuples in which a concept appears might be used as a formal definition of *understanding* of the concept.

It is now possible to sketch the broad outlines of the theory. First, information consists of linkages between properties or concepts; these atomic elements are known as *facts*. Second, geographic information is composed of a particular type of fact in which one of

the constituent concepts is geographic location (and time if relevant). Geographic location is specified in some universally understood method of spatiotemporal referencing, while the other concept in the pair is specified according to some convention that is understood by both sender and receiver. Third, the value of a fact to a receiver depends on the number of other facts already possessed by the receiver, and in which one of the constituent concepts is present; in other words, it depends on the receiver's level of understanding of the concept. These basic principles are illustrated in Figure 2.

[Figure 2 about here]

This simple framework is best illustrated with nominal concepts, but much more powerful reasoning is possible when concepts are ordinal, or interval/ratio. The elevation 8850, for example, allows naïve association with other facts about 8850, but it also allows operationalization of the concept of nearness, and reasoning about differences in height. Thus an elevation of 8840 is more similar than an elevation of 7850. Similarly since the elements of spatiotemporal location **x** are measured on interval scales it is possible to reason about nearby places, and to conduct advanced forms of spatial analysis that go far beyond the simple reasoning that is possible with nominal concepts.

## CONCLUDING COMMENTS

The simple ideas outlined in this chapter seem to address the basic requirements outlined at the start: a theory of information that is independent of medium and structure, that gives formal meaning to such concepts as fact and understanding, that is invariant under reversible transformations, and that supports the determination of value. At its base is the notion that the communication of information and the construction of knowledge occur through a process of establishment of linkages between concepts.

The chapter has assumed that linkages either exist or do not exist, and has not dealt with the possibility of partial linkage. This is the basis of uncertainty, and the subject of theories of fuzziness and rough sets (Fisher, 2000, 2001), and expressed in English through phrases such as "might be" rather than "is". Statistics often formalizes such ideas through the concept of variance, and associated confidence limits. Consider, for example, the tuple <**x**,8850±2>. One might argue that such a tuple is of less value than <**x**,8850>, but of more value than, say, <**x**,8850±100>. Moreover the theory of statistics allows inferences to be made from such statements, or from combinations of this and other uncertain statements.

How should one value such uncertain tuples? One possibility would be to assign fractional value based on the reduction of variance. For example, one could argue that without any other knowledge, the elevation of some point **x** could be anywhere in the full range of elevations exhibited by points on Earth. The uncertain tuple's value might then be expressed as $1-\sigma^2/S^2$ where $S^2$ denotes the observed variance of Earth surface elevations, and $\sigma^2$ denotes the variance in the observation recorded in the tuple.

This chapter must be understood as work in progress: much more effort will be required to develop methods that can be used as a practical basis for valuing geographic information. Hopefully the ideas outlined here provide a useful starting point.

REFERENCES


Berry, B.J.L., 1964. Approaches to regional analysis: a synthesis. *Annals of the Association of American Geographers* 54: 2-11.

Cliff, A.D., and J.K. Ord, 1981. *Spatial Processes: Models and Applications*. London: Pion.

Craig, W.J., T.M. Harris, and D. Weiner, editors, 2002. *Community Participation and Geographic Information Systems*. New York: Taylor and Francis.

Csillag, F., and S. Kabos, 1996. Hierarchical decomposition of variance with applications in environmental mapping based on satellite images. *Mathematical Geology* 28(4): 385-405.

Egenhofer, M.J., and D.M. Mark, 1995. Naïve geography. In A.U. Frank and W. Kuhn, editors, *Spatial Information Theory*. Lecture Notes in Computer Science 988. Berlin: Springer, pp. 1-15.

Fisher, P.F., 2000. Sorites paradox and vague geographies. *Fuzzy Sets and Systems* 113: 7-18.

Fisher, P.F., 2001. Alternative set theories for uncertainty in spatial information. In C.T. Hunsaker, M.F. Goodchild, M.A. Friedl, and E.J. Case, editors, *Spatial Uncertainty in Ecology*. New York: Springer, pp. 351-362.

Goodchild, M.F, 2001. Metrics of scale in remote sensing and GIS. *International Journal of Applied Earth Observation and Geoinformation* 3(2): 114-120.

Goodchild, M.F., M.J. Egenhofer, K.K. Kemp, D.M. Mark, and E.S. Sheppard, 1999. Introduction to the Varenius project. *International Journal of Geographical Information Science* 13(8): 731-746.

Gore, A., 1992. *Earth in the Balance: Ecology and the Human Spirit.* Boston: Houghton Mifflin.

Güting, R.H., and M. Schneider, 1995. Realm-based spatial data types: the ROSE algebra. *The VLDB Journal* 4(2): 243-286.

Isaaks, E.H., and R.M. Srivastava, 1989. *Applied Geostatistics*. New York: Oxford University Press.

Kuhn, W., 1997. Approaching the issue of information loss in geographic data transfers. *Geographical Systems* 4(3): 261-276.

Longley, P.A., M.F. Goodchild, D.J. Maguire, and D.W. Rhind, 2001. *Geographic Information Systems and Science*. New York: Wiley.

Marchand, B., 1972. Information theory and geography. *Geographical Analysis* 4(3): 234-257.

Onsrud, H.J, and H. Calkins, 1993. *Final Report: NCGIA Research Initiative 4: The Use and Value of Geographic Information*. Santa Barbara, Calif.: National Center for Geographic Information and Analysis.

Peuquet, D.J., 1994. It's about time: a conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers* 84(3): 441-461.

Shannon, C., 1949. *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Siegfried, T., 2000. *The Bit and the Pendulum: From Quantum Theory to M Theory, the New Physics of Information.* New York: Wiley.

Sinton, D., 1978. The inherent structure of information as a constraint to analysis: mapping thematic data as a case study. In G. Dutton, editor, *Harvard Papers on Geographic Information Systems*, Volume 6. Reading, Mass.: Addison-Wesley.

Thomas, R.W., 1981. *Information Statistics in Geography*. Concepts and Techniques in Modern Geography, No. 31. Norwich, UK: GeoBooks.

Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46(2) 234-240.

Worboys, M.F., 1995. *GIS: A Computing Perspective*. London: Taylor and Francis.

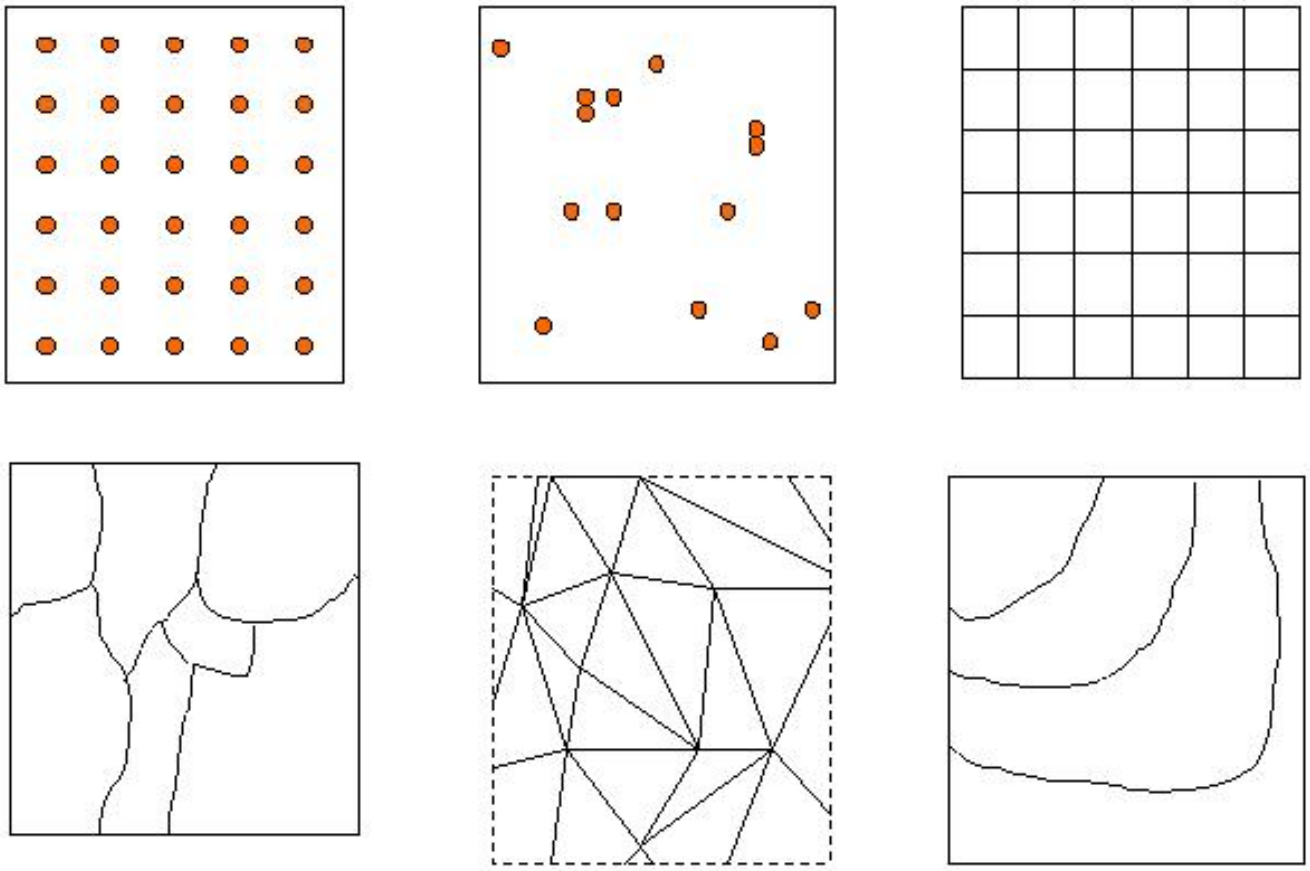

**Figure 1.** The six representations of a field commonly used in geographic databases. See text for explanation.

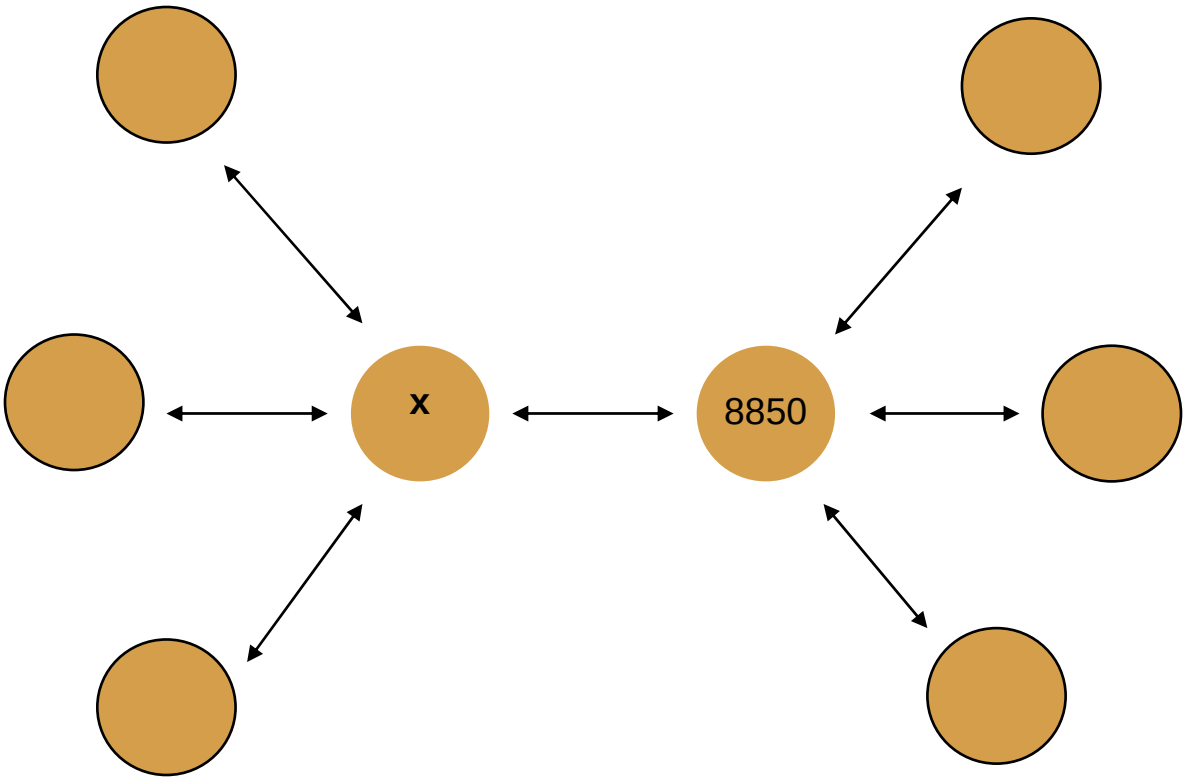


**Figure 2**. The value of a fact, represented by the connection between location **x** and elevation 8850, depends on the receiver's existing understanding of the two concepts, in the form of other facts linking these concepts to additional concepts.