The Coordinate Digit Density function and Map Information Content Analysis

Keith C. Clarke
Department of Geography/NCGIA
University of California, Santa Barbara
kclarke@geog.ucsb.edu

Sarah E Battersby
Department of Geography
University of California, Santa Barbara
batts@geog.ucsb.edu

Abstract

There are many systematic methods for encoding spatial coordinates, but there do not seem to be equivalent analytic methods for quantifying the difference in information content of these encoded data. We propose the analysis of map information content using the Coordinate Digit Density Function as a useful device for examining and comparing information quantity between spatial data in different coordinate systems (including those with non-overlapping and interwoven eastings and northings). The Coordinate Digit Density Function benefits from the fact that the coordinates used can be in any number system (decimal, alphanumeric, hexadecimal, binary, etc.), and can be used to analyze various sets of atomic data—from pairs or sets of points, to features and entire maps. Using this function and Shannon's (1948) formulae, total information content of a coordinate set can be calculated and thus compared to that of other systems or within a system. This paper presents and examines the Coordinate Digit Density Function for calculation and visual display of the degree of non-randomness in a system, and demonstrates its application in the cartographic problem of map information comparison. Specifically, we compare test data to illustrate the function's use in analysis of information reduction by map projection, reduction of precision, coordinate translation, and map generalization.

Introduction

There is a long and varied background in analytical cartography that reflects the impact of Shannon's seminal work on the theory of the flow of information (Shannon, 1948). As early as 1955, Paul Rosenberg was using Shannon's work to examine information flows in photogrammetric systems. Later considerations were by Marchand (1972), Walsh and Webber (1977), Batty (1974), Sheppard (1975) and Thomas (1981). A recent reconsideration of the value of the approach was by Tobler (1997). Generally, the prime attention given to the theory in cartography peaked in the 1970s, and has diminished since. Nevertheless, the impact of the original Shannon paper and its revisions was remarkable within computer science and communications.

Shannon's mathematical theory assigns value to the quantity of information that is communicated during an information flow. It postulates that information at the receiver end of a transmission can be reduced to knowledge about states revealed at the sender end. A communications system reduces the set of states to a finite group, for instance, the binary states 1 and 0, and then measures the probability of a message at any time sending a transmitted state to the receiver. Information flows when the message is not in the expected state. Shannon's work was involved with quantifying the rate at which

communication was possible with an imperfect communications channel. Nevertheless, the concepts of a finite state set, probabilities associated with the sets, and quantifying the minimum space necessary to represent the state set range are universally applicable. Tobler (1997), for example, considered the cases of census data for regions and gray scale levels in images. He noted the problem of state-sequence dependence. In the English language, for example, Q is most often followed by U, and T frequently by H. On maps, attribute values are always spatially autocorrelated, with structures clearly identifiable in a variogram. We postulate that while this is also true of maps and coordinates, the spatial autocorrelation inherent in the numbers and their state properties can be captured by treating the values as independent.

Slocum (1999, p. 50) reexamined Tukey's (1977) stem-and-leaf plots that seek structure in sets of numbers, primarily for application on the attribute value for maps to assist in choropleth classification. We borrow from both this and Shannon's theory, with the following assumptions. First, we assume that the information content of a map element is entirely contained in the coordinates of the features. Second, we assume that coordinates consist of a string of digits, each digit of which is in one of a finite set of known states. Third, we assume that there are no sequential state interdependencies, i.e. that each digit is independent of the remainder. We consider relaxing this assumption later. Fourth, we assume that the degree of non-randomness is equivalent to the information content. We term this information quantity, and note that information quality, redundancy and communication are not a component of the scope of this specification.

For this analysis, we propose, and illustrate by case study, a method of comparing information content between geographic data sets, called the *Coordinate Digit Density Function* (CDDF) . To illustrate the use of the CDDF function, metrics are computed for several sets of geographic data, and the results are presented for discussion. Through this analysis, it is seen that the CDDF provides a valuable method for comparison of information content between different coordinate sets.

Proposal

Let a set of geographic data consist of the atomic unit or primitive of the point *P*, represented by its coordinates (*X*). For most cartographic applications, *X* is a tuple with two numerical coordinates, though many applications require three. Some coordinate systems, for example the Military grid, reduce *X* to a single value or string, containing letters and interwoven (alternating) digits. Regardless of the number of digits or their order, we can reduce any of the tuple types to a sequence of state variables, with a finite range. So, for example, a UTM coordinate pair could be expressed in ASCII characters and the ASCII codes expressed in binary. This would create a string with *N* digits, each with two possible states. More typically, applications will be in the native digit encoding of the standardized coordinate system chosen, such as the decimal digits 0-9, decimal points, letters etc. Let each of the *N* digits in the tuple contain *S* states, where *s* can vary by coordinate digit place.

Let the expected proportion of occurrences of each state *E* for a particular digit *n* over all points *P* be uniform across states, that is:

$$E_n(s) = 1/s$$

Then any particular observed proportion of state occurrences across the same *P* points at the same digit n will vary by some amount. This amount can be attributed to each coordinate position at digit location *n*, as a value D_n:

$$D_n(s) = O_n(s) - E_n(s)$$

Summing across states gives the total digit density across states for a digit place:

$$D(n) = \sum_{1}^{s} | D_n(s) |$$

This value varies as a function with *N* values over the *N* digits in the coordinate. Note that negative contributions to the total are possible, for example when a digit is absent, and so we sum the magnitudes. Some values will be positive, when the proportion of the actual state frequencies is non-random across the *P* points. For example, the first digit of a UTM is the ordinate for the 100 km block. If all values are the same (all points are in the block), then for that digit place we have *P* occurrences of one digit out of a possible 10. For this digit, *D* will be (1.0 – 0.1) + 9 * (|0.0 – 0.1|) or 1.8. At the opposite extreme, if all digits occur at the same frequency, then *D* = 0.0. So for each decimal digit *n*, we have a value between 0 and 1.8 that corresponds to the degree of non-randomness and therefore the information quantity.

As expressed above, this value is not really independent of the neighboring digit values. For a longitude, for example, the first digit has two states, and the second ten. However, if the longitude's first digit is "1", then there are only nine second digit states, since 190 degrees is not possible. Similarly, if the first two digits are 18, then the third digit can only be 0. While Shannon did consider these dependencies, particularly with the intent of reducing their data volume, in the case of coordinates it is the spatial co-dependence that we seek to measure. Accordingly, we use only one simple value *S* for each *n*, given by the total possible number of states. One advantage of this over expedience is that for any point set *P*, we have an absolute metric for *D.*

We define the Coordinate Digit Density Function as a graphical representation of the state-derived information content at each coordinate digit expressed in the order in which they occur in the actual coordinate. Similarly, we can sum (*S*) the content across digits to get a total information content for the *P* points in the set:

$$S(P) = \sum_{1}^{n} D_n$$

This metric permits the calculation of total information content for point sets, features, lines, polygons, areas of interest, data sets, themes, and entire maps or data sets. If applied locally, that is, say as a moving computation along a line applied to a point and the two or three adjacent points forward and back along the line, it allows the local information content to be measured and symbolized. It can be used to detect data rich and information poor datasets when used in a ratio with total data size in bytes. It permits a sort of signal to noise ratio to be applied to spatial data. It also scales, that is, we can compute the relation between accuracy, generality, and geographic scale and the quantity of information contained.

Application

We applied the CDDF to polygon data from the United States and to line data from Belize and Guatamala to analyze differences in information content due to changes in projection and coordinate system, the effects of line generalization, and precision of coordinates. The US outline was that supplied by ERSI as part of the ArcView 3.2 GIS software. The Belize/Guatemala were collected by the first author using a Garmin e-map GPS receiver driven in a truck between Flores, Guatemala and San Ignacio, Belize in June of 2000. In both cases the data were originally in decimal degree format

1. *Information Content and Map Projections*

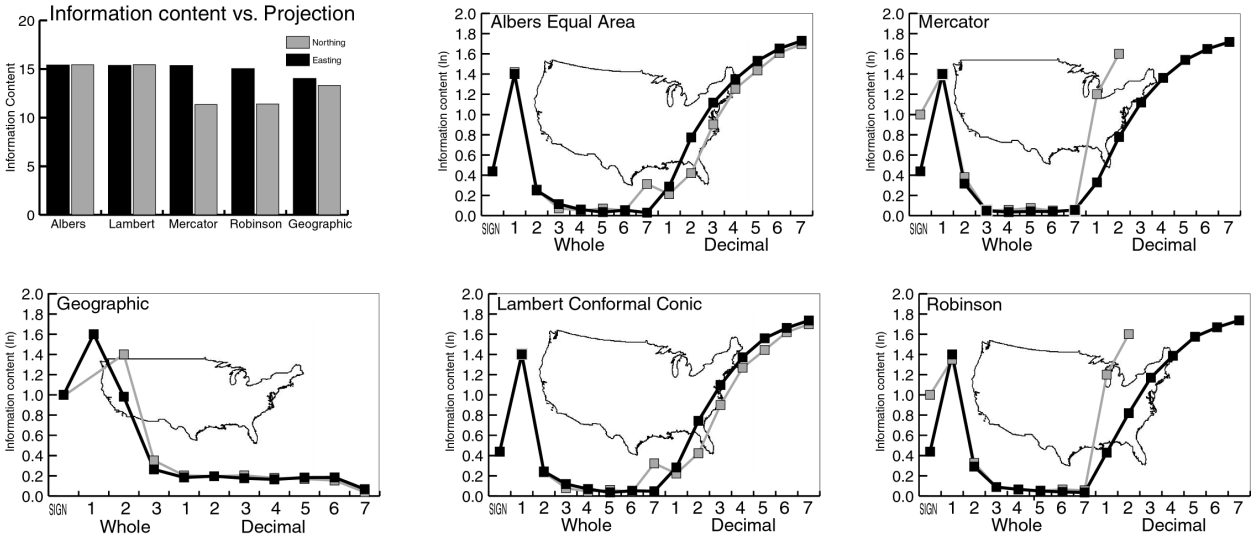
The choice of map projection is known to have a significant impact on geographic data analysis, but what is the difference in information content between individual projections and overall type of projection for the same data set of *P* points when projected into different projections?

Projections can be classified as conformal or equivalent, depending on whether shape, area or neither property is preserved in the transformation. We selected various projections belonging to each category: Lambert Conformal Conic and Mercator (conformal), Albers Equal Area (equivalent), Robinson (compromise), and compared the projected files to “Geographic” (unprojected).

Overall information content for eastings was slightly greater for all projected data than for the unprojected coordinates, with the exception of Lambert and Albers – where values for eastings and northings were roughly the same. Easting values for the differing projections were roughly the same (Figure 1). There was noticeable variation in the information content for northings, however. Albers and Lambert show little difference, and Mercator and Robinson show noticeable drops in information content. The difference in information content between eastings and northings is particularly interesting in the Mercator and Robinson data, with an average difference in information content of four points. Clearly, for conformal projections the information content does not depend on direction, so eastings and northings have similar functions and information content. This is not the case for the equivalent projections, where to preserve area the northings have been adjusted. Each CDDF, with the exception of the Geographic “projection,” bears a similar shape, however, reflecting initial non-

randomness in the sign place and the first two digits, then becoming random across the digits to the left of the decimal place and non-random to the right of the decimal place. The differences between the projections is most reflected in the single digit to the left of the decimal, then increasing into the decimal places and especially showing easting/northing differences. While much of this is non-random rounding and algorithmic error, nevertheless it is clear that for map projections "the devil is in the details."

Figure 1: Information Content For Selected Map Projections



2. Coordinate Systems

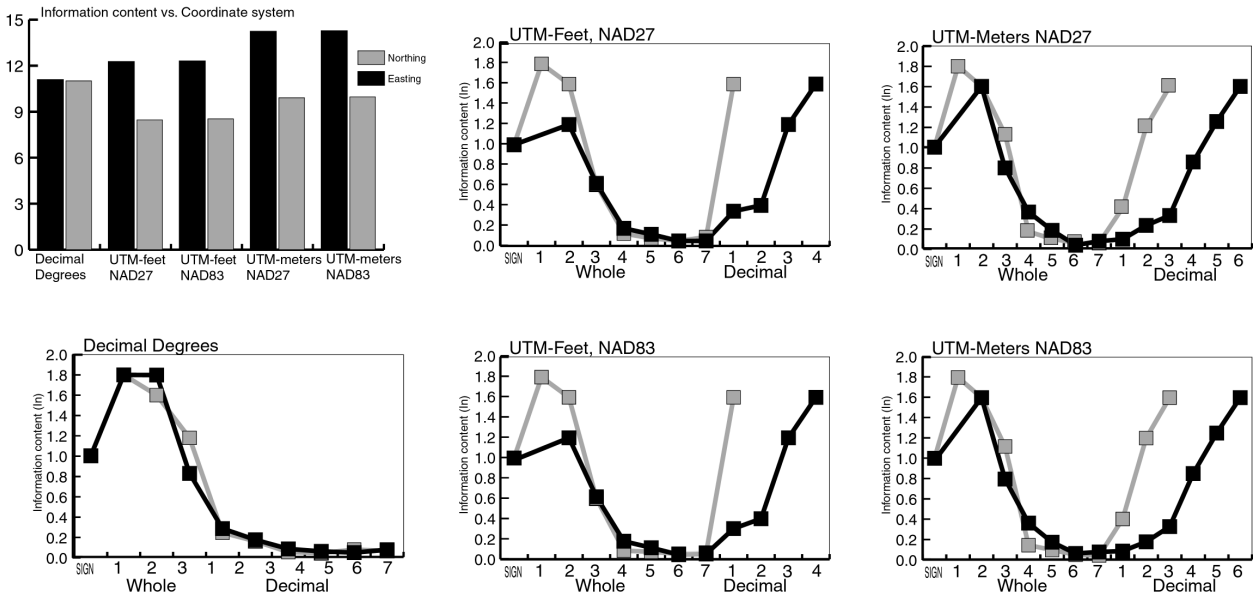
In addition to projections, we were also interested in examining changes in information content of various ways of encoding the coordinates. To do this, we compared the Belize & Guatamala data set in geographic and UTM coordinates. UTM coordinates were also calculated using different datums (NAD83-Mexico and Central America and NAD27-Central America), and units (feet and meters).

In addition to projections, we were also interested in examining changes in information content of various ways of encoding the coordinates. To do this, we compared the Belize & Guatamala data set in geographic and UTM coordinates. UTM coordinates were also calculated using different datums (NAD83-Mexico and Central America and NAD27-Central America), and units (feet and meters).

This set of data is particularly interesting in that the eastings and northings for the coordinates in decimal degrees show only a few hundredths of difference in total information content, but with change to UTM coordinates there is a sizeable difference between the two (figure 2). As the data runs primarily East/West, it is understandable that the eastings would have higher information content than the northings. While UTM is conformal, the data collected were close to a zone boundary, increasing the distortion in the northings more than the eastings. The UTMs show a more characteristic CDDF of a high at the left and right ends and a low in the digits just before the decimal point. The

information content of metric UTMs is probably due to the effect of the larger basic unit, combined with rounding. The change of datum makes no measurable difference between the information content of the coordinates.

Figure 2: Information Content for Lines in Different Coordinate Systems



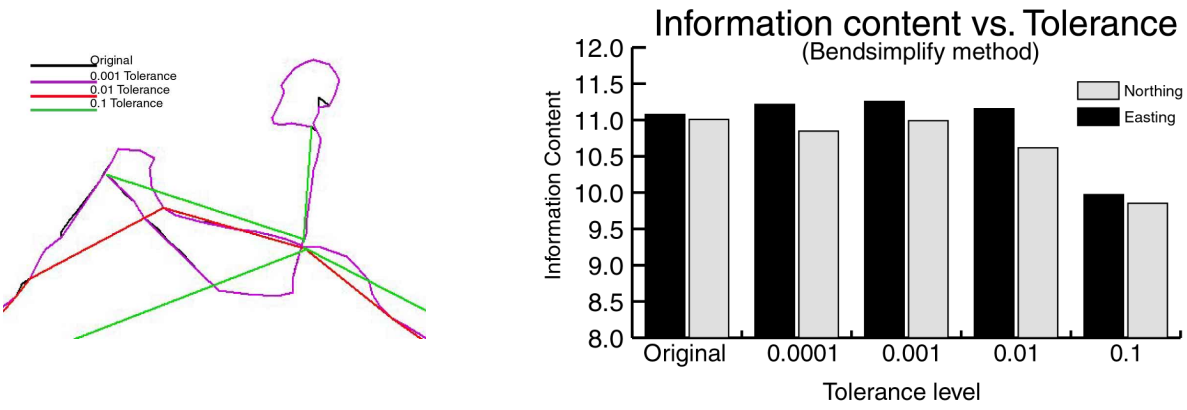
3. Generalization

Generalization is a process to remove extraneous detail, while still maintaining the characteristics of a line. The removal of the extraneous data should prove to increase the information content through reduction of redundancy in the coordinate set. This was tested using the two ArcInfo generalize options – bendsimplify and "pointremove" (Douglas-Peucker) to simplify the Belize & Guatamala dataset. The bendsimplify algorithm uses shape recognition to maintain the original line shape as accurately as possible, while still removing extraneous detail. The Douglas-Peucker algorithm is designed to simplify through retention of "critical points" that define the essential shape of the line and the systematic removal of non-critical points.

To compare different levels of generalization, we tested the data using a variety of tolerance levels. As tolerance is measured in coverage units, we opted to use values ranging from 0.0001 decimal degrees to 0.1 decimal degrees (the maximum simplification possible with our dataset).

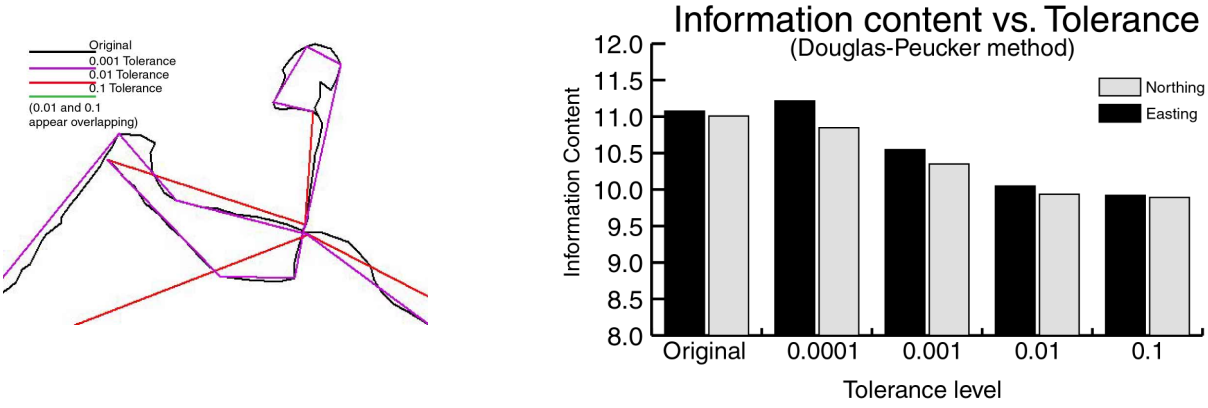
Removal of redundancy in the data should increase the information content in the dataset. We found this to be true using the bendsimplify algorithm, with a slight increase in information content of both northings and eastings between no generalization and generalization with a tolerance of 0.001 (figure 3). Information content peaked at a generalization of 0.001 and then began a slight decrease in information content at a tolerance of 0.01, and a more drastic decrease at 0.1.

Figure 3: Bendsimplify Line Generalization



Douglas-Peucker showed a similar trend in the eastings, but the increase in information content was restricted to generalization only at a tolerance of 0.0001 (figure 4). After that point, all further generalization caused a decrease in the total information content of the data. Note that the data volume was, of course, reduced in each case, as the point set *P* fell in size. The northings showed a constant decrease in information content at all levels of generalization.

Figure 4: Douglas-Peucker Line Generalization



As was predicted, both methods of generalization were successful in increasing information content of our line data set. The methods differed, however, in the amount of information gained with each generalization. Douglas-Peucker showed a greater overall increase from non-generalized to generalized using a 0.0001 tolerance, however the information content began to drop sooner than with the bendsimplify method. Using bendsimplify, the information content continued to increase through the 0.001 tolerance simplification, then began to decline, with a sharp drop between 0.01 and 0.1.

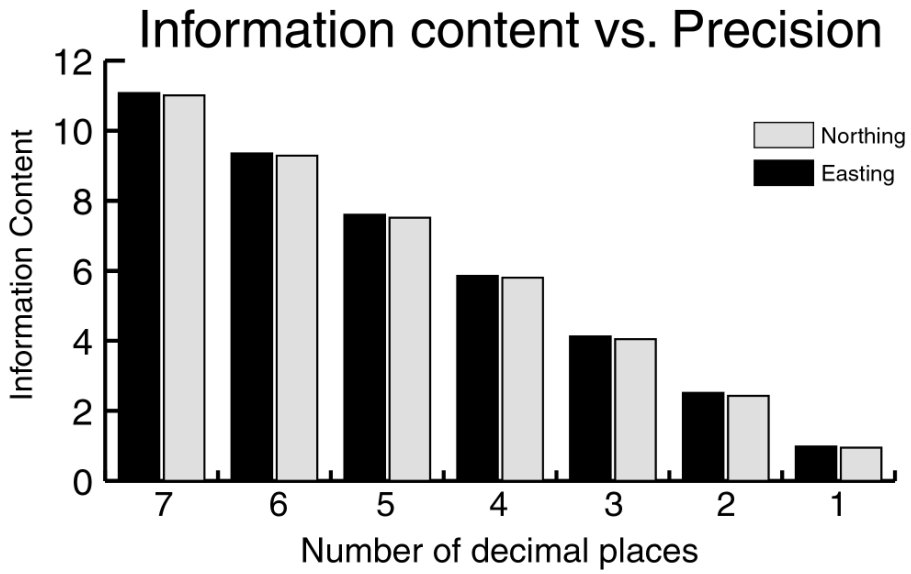
Though the level and duration of information content increase varied between the two methods, the overall change in information content – from no generalization, to the maximum level possible for the data – was roughly the same. The CDDF can identify the point at which further line generalization is not productive, as it serves only to reduce volume and not improve the information content. Clearly the bendsimplify method is effective for this data set up to the 0.001 tolerance level, while the Douglas-Peucker method peaks in content at 0.0001. In both cases, the data sets contain *more* information in generalized form that the original data sets. This amplifies the information versus data and signal to noise aspect that the CDDF captures, even in aggregate.

4. Precision

Due to problems such as rounding error, any calculation done with geographic coordinates (projection, generalization, etc.) will lead to a loss of precision in the dataset – but how does this loss of precision impact the information content of the data? As the data was captured in the geographic coordinate system, we used this system for the analysis rather than re-projecting it and losing (unintentional) additional precision.

To answer this question, we started with analysis of the Belize & Guatamala data at varying numbers of significant digits – from seven to one. As would be expected, at seven significant digits the data shows to be relatively information rich – with an S of 11.07 for the eastings and 11.007 for the northings.. The information content tapers down as the number of significant digits decreases (figure 5), showing one significant digit as being relatively information poor and highly redundant – an S of 0.98 for the eastings and 0.95 for northings.

Figure 5: Precision and Information Content



A strong linear pattern is noticeable in the decrease of information content as the number of decimal places was reduced. We calculated a linear regression for the data and found a good fit with $y = 1.694x - 0.8483$ for the eastings (r^2 0.9995) and $y = 1.6916x - 0.9053$ (r^2 0.9993) for the northings. Further analysis of other datasets showed a consistent trend of a slope of approximately 1.68 for eastings and 1.67 for northings – implying that, on average, every decimal place of precision for data in decimal degrees contributes approximately 1.7 to the final information content value.

The precision relations are strong enough to make a conjecture that the slope and intercepts are useful summaries of the CDDF. For example, the intercept is an information content of zero digits, and might be thought of as an inherent information richness or quality measure for the system as a whole. Similarly, the gradient is an "effectiveness" of adding more digits to the precision.

Further research

In work under way, we are investigating additional aspects of the CDDF. For example, localized computation within a point set allows lines and polygons to be measured and symbolized according to their information content. We plan maps of local information content for points, lines and polygons. Application of the CDDF to the attributes of images is also possible. We will conduct more data analysis to discover broader trends, since our investigations here are based on simple data sets. Furthermore, we intend to pursue further Shannon-type extensions of theory, for example, deriving full state trees for individual coordinates.

Conclusion

We have proposed and illustrated a new function, the Coordinate Digit Density Function, that is loosely based on Shannon's mathematical theory of information. The value is simply computed, and we have implemented a computer program in C that calculates the value from standard point sets. Application of the function has shown that it can demonstrate the changes in information content that follow four different cartographic transformations common in GIS work, that is map projection, change of coordinate system, line generalization, and rounding. In each case, we were able not only to quantify the amount of information, but to examine differences by digit, and between easting and northings. The CDDF seems to have numerous practical and theoretical potential for analytical cartography. For example, the application showed one way to objectively assess how much line generalization to apply to a given data set. We suspect there are many other fruitful applications of the function.

References

ESRI I (2000) "Map generalization in GIS: Practical solutions with workstation ArcInfo software", Technical Paper http://www.esri.com/library/whitepapers/pdfs/generalization.pdf

Rosenberg, P. (1955) “Information theory and electronic photogrammetry”, Photogrammetric Engineering, vol. 21, no. 4, pp. 543-55.

Shannon, C. E. (1948) “A mathematical theory of communication”, Bell System Technical Journal, vol 27, pp.379-423, 623-656.

Sheppard, E. (1975) Entropy in geography. Discussion paper # 18, Department of Geography, University of Toronto.

Slocum, T. A. (1999) Thematic cartography and visualization. Upper Saddle River, NJ: Prentice Hall.

Thomas, R. (1981) Information statistics in geography, CATMOG #31, Norwich.

Tobler, W. R. (1997) “ Introductory comments on information theory and cartography,” Cartographic Perspectives, No. 27, pp. 4-7.