Formalizing Importance: Parameters for Settlement Selection from a Geographic Database*

Douglas M. Flewelling and Max J. Egenhofer

National Center for Geographic Information and Analysis
and
Department of Surveying Engineering
University of Maine
Boardman Hall
Orono, ME 04469-5711, U.S.A.
{dougf, max}@grouse.umesve.maine.edu

Abstract

This paper describes a model for selecting features from a geographic database to be displayed on a computer generated map display. An engineering approach is used to generate a set of geographic features similar to what would be chosen by a human, without attempting to replicate the human selection process. Selection is a process of choosing from a set of objects according to some ordering scheme. Humans have a highly developed ability to order sets of things. The model presented capitalizes on this ability by relying on user-defined ordering functions to produce an ordered set (*ranked list*) of geographic features. It is possible to process systematically the ranked list such that measurable qualities of the set such as subset relationships, pattern, dispersion, density, or importance are preserved. The resulting set of candidate features is placed on the map display using accepted generalization techniques.

Introduction

Geographic databases are usually too large to be presented in their entirety to a user—displays become too crowded to identify detail or pattern. Instead, a common approach is to equip a geographic database with a spatial query language, which allows a user to specify the data of interest. While this is appropriate when requesting specific data for which a user can provide some initial information such as attribute values or location ranges, it is very cumbersome when retrieving the data for a map-like presentation. To make a "good" selection, a user needs extensive knowledge about geography—what is important enough to be displayed—and cartography—how much can be displayed on a map. Therefore, methods are needed that select a representative set of data. Any such selection method has to preserve certain properties among the features in geographic space. For instance, taking geographic features randomly for a map would be unacceptable as the resulting map could not convey the characteristic geographic properties to the users.

A major factor contributing to the interest in the *intelligent selection of information* is the attempt to automate as much as possible the process of creating maps and cartographic

* This work was partially supported through the NCGIA by NSF grant No. SES-8810917. Additionally, Max Egenhofer's work is also supported by NSF grant No. IRI-9309230, a grant from Intergraph Corporation, and a University of Maine Summer Faculty Research Grant. Some of the ideas were refined while on a leave of absence at the Università di L'Aquila, Italy, partially supported by the Italian National Council of Research (CNR) under grant No. 92.01574.PF69.

products. Automation is desirable because it would shorten the time between getting from the raw material (the observations) to the end product (the map). For geographic information systems (GISs), where the resulting maps will be short-lived, this automation has a second perspective: map-like presentations have to be created quickly and often in an *ad-hoc* manner ("on the fly") the data stored. These data serve multiple purposes and are not tailored for any particular output representation. While maintaining the original purpose of the geographic inquiry, users frequently change "scale" through zooming, or select a different map extent through panning. Several factors influence such presentations: (1) parameters, such as the real estate and its resolution available to present the geographic information, constrain how much data to select; (2) parameters, such as the relative density in cartographic space, determine when placement conflicts arise; and (3) parameters that specify the *importance* of features and, therefore, govern what data to select. This paper deals with the latter issue.

Exploration of map displays by interactively zooming needs appropriate methods for selecting subsets of features to display from a geographic database. For the purpose of this paper, a geographic database is a model of reality. It contains information about geographic features that were collected without regard to the constraints of the maps in which the features will be displayed. This makes a geographic database different from a cartographic database, which represents a model of the document, the map, and includes all details about the rendering of the information in a particular cartographic space (Frank 1991). A geographic database is therefore likely to contain many more features than can be shown on any one map.

This paper contributes to a framework for engineering a software system that generates interactive map displays from a geographic database. Ideally, such map displays would match exactly the craftswork of a human cartographer in the content, geometry, and layout; however, human cartographers' products vary quite a bit, much like text covering the same topic written by different authors. No two human cartographers would produce exactly the same map (Monmonier 1991). Therefore, the yardstick for assessing the quality of a map display will be whether or not certain spatial properties have been preserved. The overall objective is to generate a satisfactory map within the limits of an interactive computer system. To meet the requirements of the system described above, the amount of data that is actually displayed has to be reduced. A selection from a geographic database must be made in such a way that it contains the features a human would have selected as candidates to be placed on a map. As a secondary constraint for an automated system, the selection mechanism should not require extensive processing to refine the selection as the features are placed into cartographic space.

The elements on which this paper focuses are *settlements*. Settlements are critical information on any map, because map readers frequently use them as references to orient themselves and because their distribution and density describe important geographic characteristics of a region. Settlements also have the advantage of being commonly represented as points on small and intermediate scale maps (Muller 1990). A point cannot spatially conflict with itself, therefore we are able to ignore most of issues of individual feature generalization and concentrate on sets of features.

The remainder of this paper first introduces the model used in this paper for settlement selection and then discusses settlements and their properties. This is followed by a discussion of methods for settlement ranking which presents a formal approach for define ranking functions. Finally, conclusions are presented with areas for further research.

A Model for Automating Settlement Selection

In their framework for generalization, McMaster and Shea (1992) place selection just prior to and outside of generalization, as do many other researchers in automated cartography (Kadmon 1972; Langran and Poiker 1986). McMaster and Shea go on to state that after an initial selection has been made, the set may have to be further reduced. So while "gross" selection is not cartographic generalization, "refinement" of the set of candidate features by removing features from the selection is cartographic generalization. These two different selections are characteristic of the difference between the operation on geographic space that produces the original subset, and the operation in cartographic space that acts on the subset due to the constraints of the map display.

In order to make a non-random selection of elements from a set, selection criteria have to be established according to which an element will be chosen over another. An *importance attribute* has been suggested as the primary attribute in the "Competition for Space" paradigm of map generalization (Mark 1990). Under this paradigm, all cartographic symbols compete for the limited real estate available on a map. Symbols that conflict because they occupy the same map area or are too close to each other to be distinguished, undergo an examination procedure in which the one with the highest importance or certainty factor is placed on the map, while the others are dropped. The same idea underlies most name placement techniques (Freeman and Ahn 1984; Langran and Poiker 1986; Mower 1986; Jones and Cook 1989), the most advanced sub-domain of automated map generalization.

Our model for transforming features from geographic into cartographic space builds on this notion of importance as the principal mechanism to preserve the semantics of geographic objects. In this transformation of geographic features from a geographic database to a map display in cartographic space, we identify three steps: (1) feature evaluation (or ranking), (2) feature selection, and (3) feature placement (Figure 1).

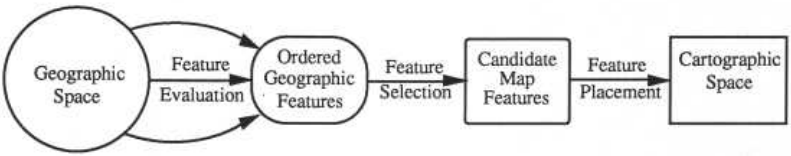

Figure 1: Transformation of features from geographic to cartographic space.

- *Feature evaluation* is the process of formally assessing the importance of each settlement in a geographic database. It generates for each settlement an *importance value* or *rank*. This procedure can be relatively simple, such as using the population of a settlement, or may involve several parameters such as a weighted balance between population, economic factors, connectivity, and availability of administrative facilities. After the evaluation is complete, the geographic semantics of the settlement have been encapsulated in the rank. The subsequent competition of map space is only based on the rank values of the settlements and their spatial locations.

  The process of ordering a set of settlements requires knowledge of the semantics of the attributes measured for the feature *and* an understanding of what information is required for the geographic problem being addressed. Defining the ranking for any set of settlements in any particular situation requires intelligence. Once the set is ordered, however, it is possible to process the set algorithmically to generate results that closely resemble selections made by human cartographers. Answering any geographic question

requires structuring of the available data with an ordering function. Ideally the ordering function should produce a single unique value for every member of the set.

- *Feature selection* extracts candidate settlements from the ranked list. There are several possible methods of selection depending on the properties that the user wants to preserve, similar to the manner in which cartographic projects are designed to preserve one or more spatial characteristics important to a map (Tobler 1979). Several of these methods have been discussed before (Flewelling 1993).

  In this model, it is assumed that not all of the settlements in the database can physically fit on the map. The process of selecting from the ordered set is on the surface simple, but an examination of published maps indicates that other criteria may be being used. There are several other factors, such as visual density, pattern, or dispersion (Unwin 1981) that geographers may wish to preserve when choosing from a set of features. Ideally, these factors should be preserved regardless of the scale of the display.

- Finally, *feature placement* introduces candidate settlements in rank order and resolves spatial conflicts in the cartographic space. It is assumed that the settlements will be represented as point features with labels. As such there are relatively few generalization operations that can be used (McMaster and Shea 1992). If settlements are placed in rank order and there is spatial conflict on the map, a settlement can be either be displaced or not placed. This is similar to name placement schemes, which usually operate as iterative processes. Successive attempts are made to find a suitable label position; only as a last resort is the feature not placed.

The particular advantage of this three-step model is that once a set of features is ranked, it is possible to produce different scale maps without addressing geographic issues for each new map display. The cartographic rules can take into consideration such criteria as screen size, display scale, display resolution, color depth, printed scale, and output media quality. None of these issues affect the underlying semantics of the features, rather they address the requirements for a legible map.

Properties of Settlement Attributes

Settlements have both spatial and descriptive characteristics, both of which may be used to determine different aspects of importance. In our model of settlement selection, the descriptive information related to each settlement in the set, such as population or number of hospitals, governs the ranking of settlements. Spatial attributes such as location and shape of a single settlement contribute to generalization at later stages of map construction.

While an exhaustive list of the descriptive parameters that can be measured for a settlement is impractical, it is possible to consider the classes of values that might be collected and determine their utility in generating an ordering for the settlements. At the simplest level, it is possible to record the presence or absence of a characteristic. For example, a city is or is not a capital. This produces a dichotomy from which it is possible to make some judgment about a set of features, but dichotomies do not have any inherent ordering. The ordering imposed on the dichotomy is simply a decision that either the presence or absence of the attribute is more important. After this decision has been made no further distinction exists among the elements of both groups.

Binary data are difficult to use with a scale-based selection threshold. There will be a range of scales where there are too few and another where there are too many features on the map. As soon as there are too many features, the only option available is to no longer show any of the feature in the set. For example, at a scale of 1:1,000,000 it is possible to show

every county seat in the United States, however at 1:10,000,000 it is no longer possible to show them all. Without additional criteria it is impossible to select from among the capitals to determine which county seats will be removed from the set. Such a threshold may have its uses in defining acceptable scale ranges for particular classes of features.

In order to be rankable, settlement attributes must have particular properties. The most common categorization of attribute data is the classical distinction of nominal, ordinal, interval, and ratio (Stevens 1946).

Nominal

Nominal data allows for the comparison of "equality" of two values, but not more. Settlement names are an example of nominal data. The name of a settlement is simply a tag to identify the place as being different from some other place. While names are important properties to identify, this piece of information is of little use in generating a complete ordering over a set of settlements. It is possible to classify settlements in terms of the relative sizes with terms such as metropolis, city, town, village or hamlet. This partial ordering of nominal values makes it possible to perform a limited amount of selection by choosing the classes of settlement to be displayed. For instance, Christaller's (1966) central place theory provides a means to classify settlements into a partially ordered set. The geographic space is divided into regions where a particular settlement is the economic center for a number of subordinate settlements. A hierarchy is developed where there are first-order settlements serving second order settlement which in turn serve third-order settlements, and so on. The result is partially ordered because it is impossible to determine whether a second-order settlement in region A is greater in rank than a second-order settlement in Region B. In this particular case there is a much lower chance that this will cause a problems with spatial conflicts because the entire space is hierarchically partitioned. This is also the case with most administrative regions, but not the case with features routes. While routes might be classified into a partially ordered set (e.g. Interstates, U.S. Routes), they often share geographic space, or due to the realities of geographic terrain, are in spatial conflict in the cartographic space.

Ordinal Data

Ordinal data establish a sequence of the elements, S_i, through an order relation, r, which is reflexive, antisymmetric, and transitive.

$$S_i \, r \, S_i \tag{1a}$$

$$i \neq j: \quad S_i \, r \, S_j \;\Rightarrow\; \neg(S_j \, r \, S_i) \tag{1b}$$

$$(S_i \, r \, S_j) \wedge (S_j \, r \, S_k) \;\Rightarrow\; S_i \, r \, S_k \tag{1c}$$

While one can determine whether one objects A comes before another object B in this sequence, it is unknown *how much* there is between A and B; however, this is not a problem in ranked lists where all that is necessary to be known is the sequence of the elements. It has been assumed in this model that the geographic semantics are encapsulated in the order.

Interval and Ratio Data

For the purposes of generating an ordered set of settlements there are no differences between interval and ratio data. Both kinds of data are true metrics since the intervals

between values are constant. The only difference being whether the zero point is set arbitrarily (interval) or not (ratio). In either case it is quite simple to determine whether the population, for example, of settlement A is greater than settlement B. For the purposes of selecting from a set of features the greater flexibility of interval and ratio data is not required. The power of these kinds of data is in the much more complex ways in which they may be combined and analyzed to produce an ordering function. For instance, Kadmon's (1972) work defines a complex set of weights and values that he used to order a set of settlements in Israel. After the set was ordered a decision was made about how many settlements where to appear on the map and the top n settlements were selected.

Ranking Settlements

Ranking a set of settlements is a transformation from a relatively information rich environment into a simple and constrained environment. In the geographic database there can be any number of parameters measured for an individual settlement. Choosing the correct function that manipulates these parameters requires cognition and understanding of the problem being addressed. For instance, in his view of the world Mark Jefferson (1917) concluded that population alone was all that was necessary to separate the important settlements from the trivial. Other techniques for ordering settlements such as those used in central place theory (Christaller 1966; Lösch 1954; von Thünen 1966) require examination of multiple variables in more complex ways.

Whatever the actual ordering function used, it is possible to state that given a set of settlements S and an ordering function o which uses the parameter values recorded for a settlement, there is a transformation function f that can produce a ranked list of settlements R:

$$f(S, o(\ldots)) \rightarrow R \tag{2}$$

such that both sets have the same cardinality (#), i.e., # (S) = # (R).

When one considers the types of information gathered about settlements, such as population, it is clear that it is impracticable to require unique ranks from this function. The probability of two settlements having equal parameter values is almost certain in a geographic database of even moderate size (Figure 2). For instance, two settlements can have the same population. Although two settlements cannot share the same location, there are cases (such as Rome and Vatican City) where two settlements are co-located when they are represented as points. It is also possible for settlements to be in the same location, but at different times. Therefore, it is necessary to develop a means for handling *complex ranks* in a partially ordered list, where a particular rank value may be attached to more than one element of the ranked list. In such cases, a decision must be made to use either selection by classification or to permit first-in resolution of conflicts.

It is possible to combine different rankings with different weights to produce a new ranking that considers multiple parameters. This approximates a more complex ordering function in f. For example:

S = Set of all Settlements
R_i = Ranked List of Settlements
W_i = Weighting Factor

$f(S, Population\ ()) \rightarrow R_{pop}$
$f(S, Employment\ ()) \rightarrow R_{emp}$
rank $(W_1 \cdot R_{pop} + W_2 \cdot R_{emp}) = f(S, Pop\_Emp\ ())$

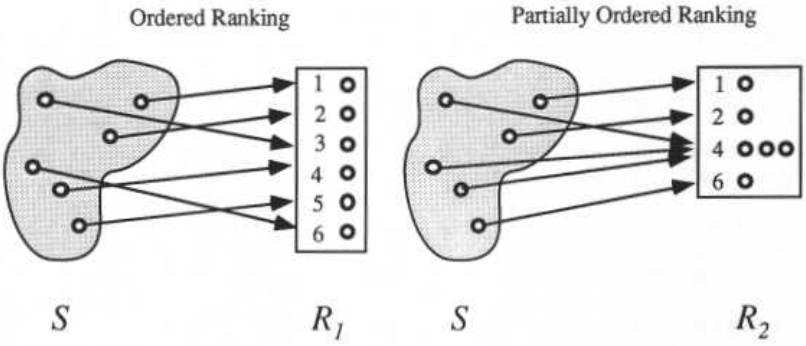

Figure 2: Transformation from an unordered set onto (a) an ordered and (b) a partially ordered ranked list.

Within the constraints of the engineering approach presented here, it may be possible to generate settlement selections with weighted rankings that are very similar to rankings created by a more complex ordering function (Figure 3). The practice is well accepted in applied and academic geography (Unwin 1981; Boyer and Savageau 1989). By providing a mechanism for storing "authorized" rankings prepared by application experts, domain specific knowledge could be encapsulated for use by non-experts.

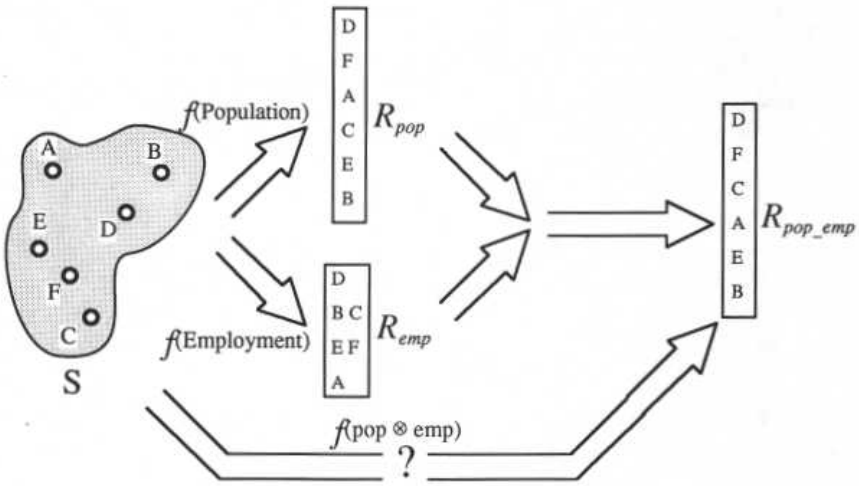

Figure 3: Combining simple rankings to approximate complex rankings.

Conclusions and Future Work

This paper has presented a model for automated map construction for an interactive GIS. The system is constrained by a requirement for frequent and rapid reconstruction of maps as scale and spatial extent change. The maps are short lived, so a reasonable approximation of the feature sets that would have chosen by a human is acceptable, as long as the most important features are shown. Settlements were chosen as the feature set to test this model.

It was shown that selecting a subset of settlements for a particular map relies on there being an order placed on the set of settlements. A random selection from the set does not preserve the geographic relationships desired for useful analysis. A framework for ordering settlements by evaluating a settlement's non-spatial parameters and ranking those evaluations has been described. Ideally the result should be fully ordered, but in practice this is difficult to achieve. Therefore, methods for processing the partially ordered list are necessary.

In this paper, no strategies for processing the ranked list were discussed in detail. Mark's (1990) "Competition for Space" paradigm suggests that only a rank order processing of the list is necessary. Where there is spatial conflict, either a "compromising" generalization method (e.g. displacement) would be used or the feature would not be placed. Whether or not this will produce an acceptable map has yet to be determined. Other selection strategies that address spatial parameters may be required to generate appropriate sets. This research is currently being conducted by the authors.

The degree to which a weighted ranking approach approximates a more complex ranking function must be investigated. If the weighted ranking approach results in selections that do not vary significantly from those in published maps then the issue becomes a matter of how rapid the system can respond to user actions. The most responsive system is one that can produce enough information about the "real" geographic world on a map for the user to make a decision that is valid in the real world.

Acknowledgments

Discussions with Andrew Frank, David Mark, and Scott Freundschuh provided useful input. Thanks also to Kathleen Hornsby who helped with the preparation of the manuscript.

References

Boyer, R. and D. Savageau (1989) *Places Rated Almanac*. New York, Prentice Hall Travel.

Christaller, W. (1966) *The Central Places of Southern Germany*, Englewood Cliffs, NJ: Prentice Hall.

Flewelling, D.M. (1993) Can Cartographic Operations Be Isolated from Geographic Space? Paper presented at the *1992 Annual Convention of the Association of American Geographers*, Atlanta, GA.

Frank, A.U. (1992) Design of Cartographic Databases. in: J.-C. Muller (ed.), *Advances in Cartography*, pp. 15-44, London: Elsevier.

Freeman, H. and J. Ahn (1984) AUTONAP—an expert system for automatic name placement. In: D. Marble (ed.), *International Symposium on Spatial Data Handling*, pp. 544-569, Zurich, Switzerland.

Jefferson, M. (1917) Some Considerations on the Geographical provinces of the United States. *Annals of the Association of American Geographers* 7: 3-15.

Jones, C.B. and A. Cook (1989) Rule-Based Cartographic Name Placement with Prolog. In: E. Anderson (Ed.), *Autocarto 9*, pp. 231-240, Baltimore, MD.

Kadmon, N. (1972) Automated Selection of Settlements in Map Generalization. *The Cartographic Journal* 9(1): 93-98.

Langran, G.E. and T.K. Poiker (1986) Integration of Name Selection and Name Placement. In: D.F. Marble (Ed.), *Second International Symposium on Spatial Data Handling*, pp. 50-64, Seattle, WA.

Lösch, A. (1954) *The Economics of Location,* New Haven, CT: Yale University Press.

Mark, D.M. (1990) Competition for Map Space as a Paradigm for Automated Map Design. In: *GIS/LIS '90*, pp. 97-106. Anaheim, CA.

McMaster, R.B. and K.S. Shea (1992) *Generalization in Digital Cartography* Washington, DC: Association of American Geographers.

Monmonier, M.S. (1991) *How to Lie with Maps.* Chicago: University of Chicago Press.

Mower, J.E. (1986) Name Placement of Point Features Through Constraint Propagation. In: D.F. Marble (Ed.), *Second International Symposium on Spatial Data Handling*, pp. 65-73, Seattle, WA.

Muller, J.C. (1990) Rule Based Generalization: Potentials and Impediments. In: K. Brassel and H. Kishimoto (Eds.), *Fourth International Symposium on Spatial Data Handling*, pp. 317-334, Zurich, Switzerland.

Stevens, S.S. (1946) On the Theory of Scales of Measurement. *Science Magazine*, 103(2684): 677-680.

Tobler, W.R. (1979) A Transformational View of Cartography. *The American Cartographer* 6(2): 101-106.

Töpfer, F. and W. Pillewizer (1966) The Principles of Selection. *The Cartographic Journal* 3(1): 10-16.

Unwin, D. (1981) *Introductory Spatial Analysis*, New York: Methuen and Co.

von Thünen, J. H. (1966) *Isolated State*, London: Pergamon Press Ltd.