Journal of the Korean Geographical Society, Vol.39, No.5, 2004(804~818)

Geometrically and Topographically Consistent Map Conflation for Federal and Local Governments*

Hoseok Kang**

Geometry 및 Topology측면에서 일관성을 유지한 방법을 이용한 연방과 지방정부의 공간데이터 융합*

강 호 석**

Abstract : As spatial data resources become more abundant, the potential for conflict among them increases. Those conflicts can exist between two or many spatial datasets covering the same area and categories. Therefore, it becomes increasingly important to be able to effectively relate these spatial data sources with others then create new spatial datasets with matching geometry and topology. One extensive spatial dataset is US Census Bureau's TIGER file, which includes census tracts, block groups, and blocks. At present, however, census maps often carry information that conflicts with municipally-maintained detailed spatial information. Therefore, in order to fully utilize census maps and their valuable demographic and economic information, the locational information of the census maps must be reconciled with the more accurate municipally-maintained reference maps and imagery. This paper formulates a conceptual framework and two map models of map conflation to make geometrically and topologically consistent source maps according to the reference maps. The first model is based on the cell model of map in which a map is a cell complex consisting of 0-cells, 1-cells, and 2-cells. The second map model is based on a different set of primitive objects that remain homeomorphic even after map generalization. A new hierarchical based map conflation is also presented to be incorporated with physical, logical, and mathematical boundary and to reduce the complexity and computational load. Map conflation principles with iteration are formulated and census maps are used as a conflation example. They consist of attribute embedding, find meaning node, cartographic 0-cell match, cartographic 1-cell match, and map transformation.

Key Words : map conflation, census maps, matching geometry and topology, map models, map generalization

요약 : 공간데이터자원이 많아 질수록 그들끼리 불일치가 일어날 확률은 높아지고 있다. 이러한 불일치는 같은 지역을 커버하는 같은 종류의 공간데이터사이에서도 일어날 수 있다. 그러므로, 이런 공간데이터를 효율적으로 연결시켜 Geometry 및 Topology측면에서 일관성을 지닌 새로운 공간데이터를 생성시키는 일의 중요성은 증가 할 것이다. 이러한 공간데이터중의 하나로서 미국 인구통계국의 TIGER파일을 예로 들 수 있다. 현재 인구통계국 지도들은 지방정부의 지도 레이어들과 공간적으로 일치 하지 않기 때문에 인구적, 경제적인 많은 유용한 정보가 지방정부의 레이어들과 연계되어 공간적으로 충분히 활용되어지고 있지 않고 있다. 그러므로, 인구통계국 지도의 위치정보는 좀 더 정확한 위치정보를 가지고 있는 지방정부의 레이어들과 융합되어 Geometry 및 Topology측면에서 새로운 정보로 대체되어져야 한다. 이 논문은 참고맵을 이용하여 Geometry 및 Topology측면에서 일관성을 지닌지도를 만들기 위한 개념적인 프레임과 두가지 맵모델을 제시한다. 첫번째 모델은 셀 모델인데 맵은 0셀, 1셀, 그리고 2셀로 구성되어진다. 두번째 모델은 수학적으로 다른 원형을 가진 물체는 지도 일반화후에도 유사성을 가지고 있다는 것이다. 새롭게 제시된 계층적인 맵 융합은 물리적, 수학적, 논리적 경계에 바탕을 두고 있고 복잡성과 계산적인 부담을 감소시킬 수 있다. 반복성을 가진 맵 융합 원리는 인구통계지도를 예로하여 형성되었다. 이것들은 속성 매치, 의미있는 노드발견, 지도학적인 0-cell 매치, 지도학적인 1-cell 매치, 그리고 맵 변형으로 구성된다.

주요어 : 맵 융합, 인구통계지도, geometry 및 topology 매치, 맵 모델, 지도 일반화

* This paper is based on a part of chapter III from author's Ph. D. dissertation.

** Senior Engineer, Industry Expert Center, Samsung SDS, Korea, hoseok.kang@samsung.com

1. Introduction

As spatial data resources become more abundant, the potential for conflict among them increases. These conflicts can be described as both spatial and non-spatial and can exist between two or many spatial data sets covering the same areas and categories. Using a uniform spatial reference when creating multiple spatial datasets could prevent these conflicts to a great extent. However, using one spatial reference is almost impossible since there are so many independent agencies that create spatial layers based on their own internal needs and purposes. Therefore, it becomes more important to be able to effectively relate these spatial data sources with others and then create spatial datasets with matching geography. For example, spatial data that has more positional accuracy can be related with other spatial data that has less positional accuracy but more valuable attribute information to get a new consistent map that has more positional accuracy than the original and at the same time valuable attributes. One such spatial dataset is the US Census Bureau's TIGER file, which includes census tracts, block groups, blocks, and also roads, railroads, limited hydrology, and so on. For the remainder of this paper, the term census maps will be used for census tracts, block groups, and blocks. US Census Bureau provides basic statistics about the people and the economy to the congress, the executive branch, and the general public every 10 years (Marx, 1984). There are tremendous demands on US Census statistics that are closely related to geography but unfortunately, the lack of positional accuracy in census maps makes them less use and less effective.

Currently, census maps that usually do not match the localities' more precise mapping layers do not allow local users to correlate between them. Therefore, in order to accomplish this task, the geometry and topology of the census maps must closely match the local and more accurate sources.

Map conflation can be defined as combining two or more map data sets from different sources to form a new spatial data set with a unique representation and consistency. Automated map conflation has been implemented since Saalfeld suggested automated map conflation in which recursive point feature matches are performed (Saalfeld, 1993). Different versions of map conflation could arise in the following cases.

· When combining the same type layers of the same region.

· When combining the different type layers of the same region.

· When combining the same type layers of neighboring regions.

There are multiple reasons for the extensive difference between the geometry and topology of census maps and local government's GIS base. The first reason has to do with the scale at which each coverage was originally created. Census maps are created from 7.5', 15', and 1:100,000 scale hardcopy maps by heads-down digitizing or scanning (Callahan, Year unkown). However local governments make their own base maps by field surveys or heads-up digitizing of digital orthophotos that are very high resolution to provide parcel level information. For example, there are 3 scales of orthophotos that are at 1: 1,200, 2,400, and 4,800 in Delaware County, OH, USA. Other maps such as road centerlines and parcel boundaries are created at the same scales. When two different scale maps are overlaid at a larger scale, jagged effects(coarse representation) cannot be prevented in the smaller scale map, because the smaller-scale map becomes just larger (zoom in) than its own scale without adding any details.

The second reason relates to generalization. Since the real world is too complex for our immediate and direct understanding and there is a limited space on a hardcopy map, generalization must be applied when a paper map is created. For example, generalization operations include selection, simplification, exaggeration, classification, displacement, symbolization, and so on. It is an ill posed problem to recov-

er the original feature from the generalized one without extra information. This can be described as an inverse generalization problem(Can we invert the generalization process? If not, can we generalize a large scale map to small scale in a way that helps to match feature? If so, for which feature or area can we facilitate matching through appropriate generalization procedures?). This can also be described as a conflict between analog and digital based data. These two main factors make map conflation complex and not easily defined. Since map conflation is an ill posed problem as discussed above, it is extremely difficult to find fully automated solutions.

The purpose of this paper is to formulate a conceptual framework and mathematical model of map conflation in order to make geometrically and topologically consistent source maps according to the reference maps. The new approaches such as hierarchical conflation, conflation of heterogeneous spatial data types, and 0 and 1 cell match methods are also discussed. This will provide a unique solution of map conflation for local governments. Census maps and local government' GIS base(Delaware County, OH, USA) are used as an example.

2. Mathematical Models of Maps

Two distinct mathematical models are provided for map conflation. The first and more familiar model is based on a mathematical theory of cellular surfaces composed of idealized cell objects of zero, one, and two dimensions. The first model looks at graphs embedded in orientable surfaces. The second model is based on a different newly-formulated mathematical principle that requires all map features---line, point, or otherwise---have non-zero area. This newly-stated mathematical principle, in turn, derives from the physical reality of perceptual acuity that requires an object to have significant extent in all directions in order for the object to be visible and be seen.

Whereas the first model allows for collapsing of features (forcing areas to become lower-dimensional lines or points), the second model does not need. When a lake representation changes from a polygon with interior to a single 0-dimensional point, the transformation is many-to-one and, hence, not invertible. When a lake representation changes from a polygon with interior to a "fat" point (i.e., a small disk), then a local homeomorphism (a bi-continuous bijection) may still be established. This second theory helps to localize feature matching in map conflation by searching for and then establishing a one-to-one bi-continuous relation everywhere.

1) The Cell Model of Maps

In the traditional mathematical model for a map (Corbett, 1979), a map is a cell complex consisting of 0-cells, 1-cells, and 2-cells. Removing the open 2-cells from a rectangular map or a globe leaves behind the 1-skeleton, a plane graph embedded in Euclidean space. This model remains the basis for many systems currently in use by US federal government agencies, including the Census Bureau's TIGER files, the DLG files of the US Geological Survey, and NIMA's Vector Product Format (VPF) files.

For conflating datasets that use the cell model, difficulties arise when no cell-structure-preserving homeomorphism exists. For example, one cannot match maps, point for point, when a corresponding feature pair is represented on the two maps by cells of different dimensions. Such cells of different dimension admit no homeomorphisms whatsoever between them. In addition, mapmakers occasionally conceptually change dimension of the object they are depicting (a 2D road has width, a 1D road centerline does not; a 2D manhole cover has area, its 0D point location on a map does not) without changing the drawn representation. Drawn map features are not idealized mathematical objects, although they may purport to represent idealized mathematical objects (e.g., a boundary line is a zero-width line, but it cannot be drawn as such). Finally we note that decreas-

ing a map's scale requires more than a simple contraction. Mapmaking practice requires that the positional adjustment not be uniform (exaggeration may be necessary to capture a feature's character).

2) A Homeomorphism Model for Map Conflation

One idealized mathematical model for a map is a homeomorphism (a bijective bi-continuous function) or a diffeomorphism (a homeomorphism that also has partial derivatives everywhere) of the Earth's surface. Diffeomorphisms admit partial derivatives between the coordinates of one surface and coordinates of the other, which makes sense for a mathematical surface (sphere or ellipsoid) or a surface satisfying spatial relations that are described in term of potential fields and differential equations, such as the geoid. The carrier topology of a surface with a cell decomposition is the underlying point set topology of the surface. The carrier topology can be derived from any cell decomposition. Nevertheless, the carrier topology is unique, and therefore any two different cell decompositions of the same region must produce the same carrier topology.

By this model, all representations of a single region are homeomorphic (or possibly even diffeomorphic) to a common surface, and, hence, are homeomorphic (or diffeomorphic) to each other. For example, all map projections are diffeomorphic away from "interruptions," the discontinuities resulting from cuts made to flatten the map.

For two idealized maps, this homeomorphism must exist, and the homeomorphism can be explicitly described. Map conflation in this framework consists of nothing more than making explicit the existing homeomorphism. Corresponding features need to be mapped onto each other to establish a location correspondence. Features appearing on only one depiction need to be assigned an appropriate location on the other. The Jordan-Schönflies Theorem (see, for example, Mohar, 2001, for a statement of the theorem) implies that every point or feature that is interior to one or more closed simple loops must be matched with a unique point or feature within the corresponding loops. This property greatly reduces our search for matching features to within a nested collection of subregions.

Matching operations in map conflation are now defined in two ways, that is, totally homeomorphic and potentially non-homeomorphic matching.

If corresponding matching features within a subregion are of the same dimension, and hence homeomorphic, conventional matching strategy may be applied. If corresponding features are not homeomorphic, then an operation called topological surgery (replacing an entire neighborhood by a more detailed neighborhood that agrees on the boundary) may be implemented.

3. Map Conflation Principles

A spatial process diagram for map conflation is presented in Figure 1. It has the capability to produce geometrically and topologically consistent maps. Each component use mathematical and analytical theories that deal with irregular domains. Higher or general level boundaries represent outline views of geographic spaces. Details can be examined in lower level boundaries with successively finer resolutions. An attribute-embedding component was developed to assign a spatial key to each line segment, which links to source database. A meaningful node is defined not according to geometry properties, but according to attribute properties. The categories to distinguish meaningful nodes are (1) same type and (2) same junction. Cartographic 0-cell matching uses a *0-cell conflation test* to see if the neighborhood of the 0-cell has a legitimate topology, in other words, if the two subregions are homeomorphic. Geometry, topology, and attribute match tests are performed if the regions are homeomorphic. Using two matched 0-cell pairs and its correspondent 1-cell in the source map, cartographic 1-

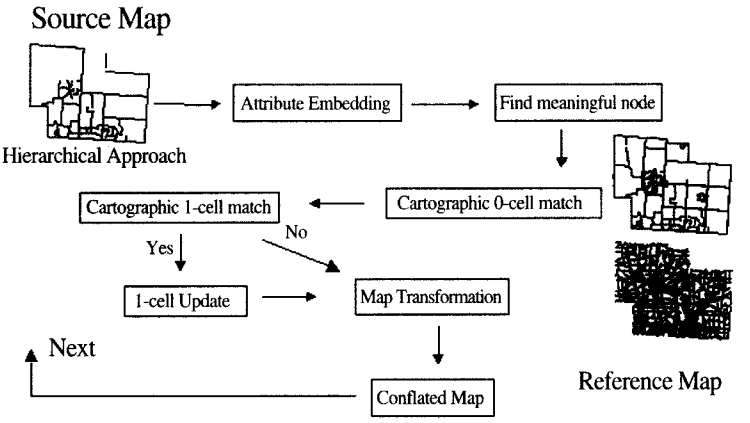


Figure 1. Spatial Processes for Map Conflation

cell match is implemented.

Finally, map transformations using local triangulation are applied to transform the features of lower boundary inside higher boundaries. The results of map transformation are used as the inputs for the next level conflation. These spatial processes may be iterated to match elements within lower boundaries. Several characteristics of the suggested map conflation are explained in the next sections.

1) Hierarchical Conflation

Hierarchical structuring has many advantages not only for GIS fields but also for many other fields as well. The main methodology of hierarchical structuring is to subdivide complex reality or to refine broad problems to provide various levels of understanding. According to (Timpf, 1998), there are three types of hierarchies (aggregation, generalization, and filter). Useful hierarchical approaches to solve GIS related problems are abundant in literature. (Car, 1996) shows that human beings use hierarchies extensively to simplify their conceptual models of reality and to solve spatially reference problems more efficiently. She builds hierarchical road structures such as expressways, highways, and local roads to find optimal paths. Hierarchical spatial reasoning is based on the cognitive assumptions (Car, 1994a, 1994b).

By the way of hierarchical structuring, complexity and processing time can be reduced because the effective problem area becomes more manageable and the chance of better performance is increased because an ordered structure is provided. The U.S. Census Bureau uses the following

hierarchical geographic entities (Figure 2). These geographic entities are defined by legal and statistical criteria.

The regions form a nested structure in which a census tract includes its block groups and a census block group includes its blocks. In order worlds, the external boundary of all block groups is the boundary of census tracts and the external boundary of all blocks is the boundary of block groups.

Table 1 shows the hierarchical relationship of cen-

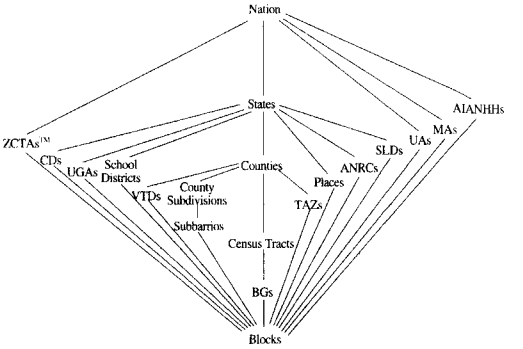


Figure 2. Hierarchical Relationship of U.S. Census Geographic Entities (Census Bureau, 2000)

Table 1. The Hierarchical Relationship of Census Tract, Block Group, and Block

	Tract	Block Group	Block
Purpose	Decennial census data	Decennial census data	Decennial census data
Area	A small, relatively permanent statistical subdivision	Contiguous area inside census tract	The smallest of the census geographic areas
Size Range	1,500 - 8,000 inhabitants, optimum size is 4,000	600 - 3,000 inhabitants, optimum size is 1,500	Average 100 inhabitants
Boundary	Visible feature, legal or governmental boundaries, non-visible in some instance, and always nest within counties	Visible and non-visible feature	All cases, visible features such as streets, roads, streams, and railroad tracks, and invisible features such as city, town, township, county limits, and short imaginary extensions of streets and roads.
Revision	Seldom, but if there is physical changes such as new highway construction. Sometimes spilt or combined if there is population changes	More than Census Tract	More than Census Block Groups
Spatial Relation	Parent	Child	Grand child

sus tract, block group, and block. Those registered boundaries are changed every 10 years based on population numbers as illustrated in Table 1. If there is considerable land development in an area, then there are even more opportunities to redefine boundaries by repartitioning and consolidation. Figure 3 shows a step in hierarchical conflation for census maps.

2) Heterogeneous Map Conflation

Heterogeneous reference maps may be used to find corresponding anchor points of source maps. Heterogeneous reference maps are polygon and polyline maps such as township boundary, road centerline, hydrographic, railroad, municipal boundary, and parcel boundary. There are also two kinds of heterogeneity. One refers to the map type itself, and the other refers to the variety of components that bound a polygon.

The boundaries is homogeneous in terms of being a polygon boundary itself, but they are not homogeneous in terms of their attributes within geography feature.

For example, for a single census tract, some line segments are roads and others are political boundaries(Figure4).

If various types of maps are overlaid, new point features, intersections between the layers, are created and new topology can be also built. As far as map conflation is concerned, a new topology build in the reference map may not be necessary because the

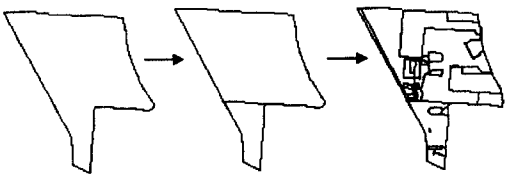


Figure 3. Hierarchical Subdivision of Census Tract, Block Groups, and Blcoks

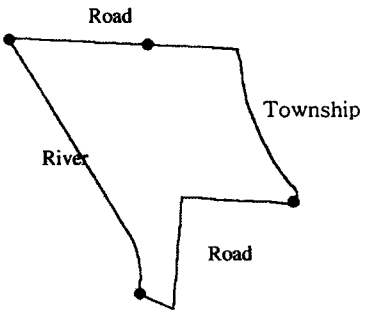


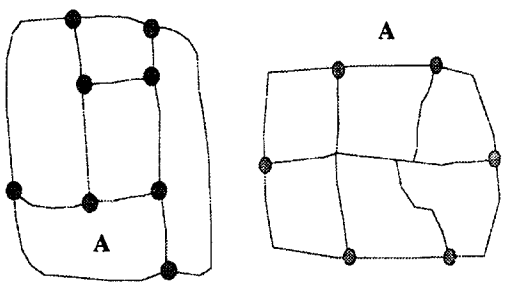
Figure 4. Heterogeneous Strings on a Polygon Boundary

new topology build may not necessarily reflect any features in the source map. Intersection points are not pre-calculated because the situation when they are needed is known in the process of matching. Therefore, this process is carried out on the fly. The advanced intersection finding algorithm, a sweep line algorithm ($O(n \log n)$), could be used but a brute force algorithm ($O(n^2)$) may also be acceptable if the number of line segments is small.

3) Anchor Nodes

In most recently developed map conflation systems, anchor nodes are defined as well distinguished points such as intersection and turning points so that we may automatically and easily distinguish them in the digital map. However, for hierarchical boundary map conflation, there are no necessary conditions for a point to be an anchor node. Since difference types of data constitute census map boundaries, it is useful to match homogeneous strings by type before finding anchor nodes (Figure 4). Besides those anchor nodes, in order to make connections between tract, block group, and block, junction nodes among them may be useful (Figure 5).

For example, junction nodes between tract polygons play a role in making a consistent topology when they are combined. Junction nodes on a tract boundary are nodes that share with the lower level features, block group and block, inside a tract polygon. These relations are still valid in the next lower level. The attribute embedding process depends on spatial data system. For example, each line segment of census map boundary is assigned a unique key (tlid) that will lead a link to the TIGER/Line Database (Broome, 1990) which, in turn, has all attribute information.



Junction nodes between tract polygons

Junction nodes on a tract boundary

Figure 5. Junction node illustration

4) Map Feature Matching

Many methods have been developed to find 0-cell matching pairs between two maps. These methods mainly depend on attribute, geometry, and topology information to find possible candidates and then statistical methods can apply to choose the best-fit pairs if there is no unique exact match case. The proposed attribute, geometry, and topology matching module finds match pairs from multiple candidate points. Attribute matching may be separated into exact and approximate matching tests. Exact matching tests checks if two maps have exact common parts in prefix, name, suffix, and other information. Approximate match adopts generalization operations. Geometry and topology matches have been implemented as follows.

0-cell

Geometry: point - point (proximity)
Geometry and Topology: node - node (neighbor proximity, degree)
Heterogeneous point (intersection of different type features)

1-cell

Geometry: string - string (distance, direction)
Geometry and Topology: chain - chain (neighborhood, connectivity)

2-cell

Geometry: G polygon - G polygon (area, perimeter, slope)
Geometry and Topology: GT polygon - GT polygon (neighborhood, connectivity)

Multiple candidates may be chosen within any

given buffer distance but it is not possible to set a universal buffer distance. Instead, the n nearest points from the location where the point of source map is projected onto reference map were chosen for investigation. Since there are several reference maps, those multiple candidates may even include intersection points between reference maps. First, 0-cell matching criteria based on proximity, degree, neighbor proximity, and neighbor degree are applied. Next 1-cell match is evaluated based on distance, direction, neighborhood, and connectivity. Finally, 2-cell matching is made based on area, perimeter, neighborhood, and connectivity, if applicable. Neighbor checking gives further evidence to use to avoid choosing a false match(Figure 6).

Without neighbor checking, node 1 may be incorrectly matched with node 2′ because they are the closest pairs and their degree (sharing number of lines on a node) and direction are matched. However, this case will be rejected with neighbor checking because neighbors (node 2 and node 3′) do not have matching node direction. Therefore, the second-closest pairs (node 1 and node 1′) will be chosen.

Since there are multiple match criteria and multiple candidates, several scenarios are possible. The easiest case to resolve is that all match criteria are passed only for the closest match pairs so that the rest of candidates are not even considered. The uncertainty for determining correct match pairs occurs when there are some candidate matching pairs for which only some of the matching tests are passed. In that case, one may try to establish weights for the matching criteria and then try to find a best fit based on those weights. This can be written as follows.

$$\text{Best fit} = max(P1 \sum_{i=1}^{n} wiMi, P2 \sum_{i=1}^{n} wiMi, P3 \sum_{i=1}^{n} wiMi, P4 \sum_{i=1}^{n} wiMi, \cdots)$$

Where:

Pn : Candidate match pairs

wi : Weight

Mi : Match test

However, assigning weights is not anything like a straightforward or simple task. For example, distance may be more important in some cases, whereas the degree of a node may be more important in others. It may turn out that no rules exist because the domain is too irregular and/or the relative weighting scheme is too subjective.

The most important thing is that modeling uncertainty is not needed if unique match candidates of the same dimension exist. In other words, we have to find as many exact match cases as possible and then model uncertainty if there is no one best choice. Therefore, if there are no exact match pairs in matching test, determining weight functions to get an approximate match would not be the best choice. Another possibility is to change geometry and topology of the source data to help derive the exact match. This can be done by employing the homeomorphic model of map conflation.

From the point view of completeness in making a map or homeomorphism, map features between maps that have same coverage should be matched each other. However, this is not case in map conflation because one map is at a smaller scale, and generalization procedures have already been conducted. Therefore, the uncertainty in any matching process already exists. It is impossible to invert the general-

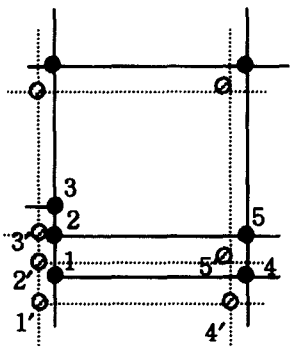


Figure 6. Neighbor Checking in Node Matching

ization process, but it may possible to recover the original shape before generalization from reference maps. The proposed method provides consistent geometry and topology for a candidate point in the source map after modeling generalization in the reference map. The advantages of this approach are as follows.

· It can find many exact match cases to reduce uncertainty.

· It can change geometry and topology in the source map to represent more realistic geometry and topology (rebuild one-to-one relation).

· It can allow one candidate point to be divided into multiple points to restore the original shape (add details).

(1) Generalization and Matching

This section discusses the relation between generalization and matching strategies and provides *0-cell conflation test*.

① Graphic and Human Visual Resolution Limitation

Cartographic generalization, an essential component in mapmaking and for successful map communication, is the process of selecting and simplifying the reality according to the scale, the objective, graphic limits, and map users of a target map(Timpf, 1998; Hake, 1975; Robinson, 1995; Anson, 1993; Brassel, 1988; Buttenfield, 1991; McMaster, 1992; Brassel, 1988; Ramirez, 1993). Among generalization factors, the scale, the objective, and map users of source maps are given in the case of census map conflation.

Census maps are created by digitizing USGS 1:100,000, USGS 7.5′ (1:24,000), and USGS 15′ paper maps. The objective of the USGS map series is to represent entire Nations in the forms of paper maps. The 1:100,000 series started earlier than 7.5′ and 15′ series. At those scales, most of the major, intermediate, and minor of road, railroad, and hydrographic boundary are represented (Thompson, 1988). There is, however, a high probability that some minor features such as private roads, drawbridges, and exposed wrecks will be missing from the 1:100,000 scale maps (Thompson, 1988). Map users of these maps are general public so that map representation and symbols should be widely understandable. Therefore, one can assume that the boundary information for census maps is all there. One can also expect that those boundaries are subjected to a cartographic generalization process, which could change geometry and topology of boundary information, because of graphic and human visual resolution limitations. For example, two lanes might have been combined into one lane if the space between two lanes is small enough. Two intersections that are close each other might have been combined (Figure 7).

There are physical distances of graphic and human visual resolution limitation for linear features. The separation distance between two lines should be at least 0.15 mm and this distance should be even greater if the lines are very fine (Cuenin, 1972; Keates, 1989). The minimum line width that is discernible is 0.06 mm. Finally, the length of a line should not be less than 0.6 mm (Cuenin, 1972; Keates, 1989).

② Generalization and 0-cell Match

For purposes of finding matching pairs for map conflation, the scale difference between source and reference maps should be taken into consideration before geometry, topology, and attribute matching

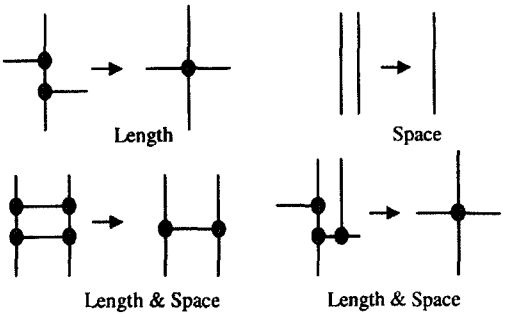


Figure 7. Generalizations of Linear Feature that Change Geometry and Topology

rules are applied because a scale decrease may result in the maps no longer being homeomorphic. Therefore, the following *0-cell conflation test* is proposed.

$$\text{graphic limit} \le k(\text{distance}\frac{\text{Map Scale in Source Map}}{\text{Map Scale in Reference Map}})$$

k is a constant that controls displacement; if there is no displacement, then k is 1. Graphic limit is 0.15 mm and 0.6 mm for the separation and the length, respectively. "Distance (L in Figure 8)" refers not to the distance between candidate match pairs in two maps but rather to the distance between a node and its neighbor point in the reference map.

In Figure 8, the map intersection at 1′ is the source map, and the solid line map with the multiple intersections at 1 and 2 is a reference map. When node 1′ looks for its corresponding node in the reference map, the *0-cell conflation test* is applied to see whether node 1′ is a potential generalized node (in other words, to check for local homeomorphism). L, the distance between node 1 and 2 in the reference map, is used to see if the rescaled distance is within graphic limit or human visual limitation. If the rescaled distance is within graphic limit, then node 1′ finds its multiple match pairs by relaxing its requirement of maintaining topology. In this case the geometry and topology of the source map should be modified according to that of the reference map.

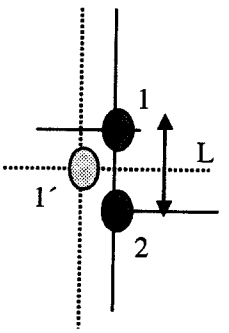


Figure 8. 0-cell conflation Test Illustrated

Figure 9 shows a topological surgery operati-on for the node that has multiple candidate matching pairs. The node 1′ of the left map of Figure 9 finds its multiple match pairs in Figure 8. The four regions (A′, B′, C′, and D′) in the source map (the middle map) are treated separately and then are replaced with their four-match regions (A, B, C, and D) in the reference map (the right map). Finally, one four-way intersection becomes two three-way intersections to reflect true topology after conflation.

This example is a good illustration of generalization by cartographic simplification. If there is generalization by feature merging, such as a divided highway shown as a single lane, then the two lanes may still be recognized using the *0-cell conflation test*. From the point view of census geography, however, it is not important to separate lanes because the median zone contains no demographic data. For example, major highways usually have a median area. These roads are nevertheless described single connector using a centerline in the small scale and two distinct connectors, one for each direction in the large scale. Therefore, as long as a globally consistent topology is maintained, either connector may be chosen to match the single connector representation.

In summary, if there is a many-to-one or a one-to-

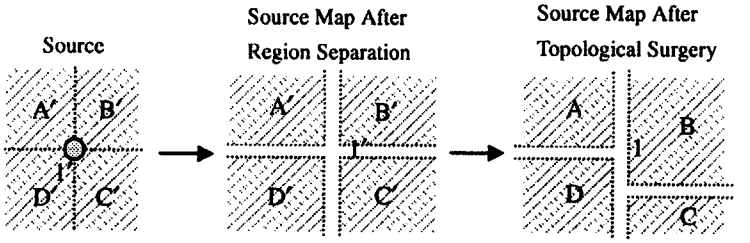


Figure 9. A Topological Surgery Operation

one matching relation between reference and source and their relation does not affect topology outside neighborhoods that have matchable boundaries, then the *0-cell conflation test* may be relaxed. For example, displacement should not alter topology after generalization. One can also argue that there should always be corresponding features in the larger scale reference map for all of the generalized linear features in a source map.

③ Generalization and 1-cell Match

In general, when two strings (polylines made up of a sequence of connected line segments) are tested for matching, the shape of the strings and their relative positions and orientations are used as criteria for selection (Hangouet, 1995; Saalfeld, 1986). Sometimes may it be satisfactory to use simple shape, location, and orientation matching rules for 1-cell matching. However, since map conflation includes consolidating maps at different scales, matching features may fail to exhibit even similar geometry. For satisfactory matching of features of map at different scales, we must understand how generalization may alter geometry. Some of the effects of generalization have been well studied. Topfer's selection law, for example, describes the relationship between the number of features and map scales (Topfer, 1966). One classic string simplification algorithm, the Douglas-Peucker algorithm, has been used to reduce the number of vertices of the string (Douglas, 1973). A pre-selected buffer-distance is used to remove a polyline's vertices if their distance from a simplifying line is less than the buffer-distance. A modified version of the Douglas-Peucker algorithm has also been developed to preserve topology in a neighborhood of the simplified polyline (Saalfeld, 1998).

The vertices of the string in the reference map are given. Therefore, string comparison may still be implemented after the line simplification of a string in the reference map.

(2) Graph Theory and Match

Suppose that two pairs of matched nodes (A↔A′, B↔B′) are found by 0-cell matching and there is a string (A′ to B′) in the reference map. The task is to find the string in the source map that corresponds to the target string (A′-B′) in the reference map. Figure 10 shows this situation.

Since the underlying 1-skeleton of a map is a plane graph, there are many graph matching algorithms that deal with aspects of this problem. There are many approaches to select a candidate string from the source map's graph if start and end nodes are known. One might be tempted to try a greedy algorithm, such as: starting from start node A, first choose a segment that is connected to node A and remains closest to the string A′ - B′. But this approach has several disadvantages.

There is no universally agreed-upon procedure for determining closeness. For example, metrics involving length or direction can be used to select the "closest" line segment. But this is case by case because one string is finer and the other is coarse. A good partial string match at the beginning segments

Figure 10. Finding Matched String After 0-cell Match

may lead to a very poor overall string match.

Locally acceptable matches may fail to capture the true overall shape of the string. Therefore, global comparison between strings should be done to consider generalization effect. However, this approach could force one to examine many candidate strings that start from node A and end at node B in the graph. Most of them are simply not worth considering. A tolerance buffer may be used to restrict the search space since the geometry of the reference string (A'-B') and its two end matching nodes(A-A' and B-B') are known.

Using only those edges within the tolerance region, the k shortest paths are found. The k shortest paths algorithm uses the well-known Dijkstra method repeatedly (Katoh, 1978). Dijkstra's method finds the shortest path between two vertices in a graph. If the graph G =(V, E), where V is a given set of vertices and E is a given set of edges, then Dijkstra's method stores the total cost from the source vertex to the current vertex by using temporary and permanent labels. The temporary labels are vertices that have not been reached and the permanent labels are given to those vertices whose cost to the source vertex is known (Gross, 1998). Dijkstra's algorithm to find the shortest path from vertex s to v is as follows (Gross, 1998).

Initialize the Dijkstra tree T as vertex s.
Initialize the set of frontier edges for tree T as empty.
dist[s]:0
Write label 0 on vertex s.
While Dijkstra tree T does not yet span G
For each frontier edge e for T
  Let x be the labeled endpoint of edge e.
  Let y be the unlabeled endpoint of edge e.
  Set P(e) = dist[x] + w(e).
  Let e be a frontier edge for T that has the smallest P-value.
  Let x be the labeled endpoint of edge e.
  Let y be the unlabeled endpoint of edge e.
  Add edge e (and vertex y) to tree T.
  dist[y] := P(e)
  Write label dist[y] on vertex y.
Return Dijkstra tree T and its vertex labels.

The time complexity of Dijkstra method is $O(v^2)$, where v is the number of vertices. The graph G (V, E) created from tolerance region may have islands and dead ends. In order to make simple graph in terms of start and end vertex, delete islands and dead ends are performed before k shortest path implementation.

The k shortest paths algorithm uses the additional constraint that additional paths branches from a specified initial portion of the first shortest path. In other words, if there is branch from the first shortest path, Dijkstra method is repeated but the start node is changed to the next branch vertex. The final algorithm to find k shortest paths for map conflation is as follows.

set tolerance region
G(V, E) is initialized
set start and end vertex
delete islands and dead ends
call Dijkstra to find 1st shortest path
while if there exists branch in the n-th path
  set new start vertex
  call Dijkstra to find n+1-st shortest path
Connect new start vertex to its parent

Finally, line generalization on those candidate strings may be implemented to see how a large-scale string can be modified to produce a representative small-scale string. 1-cell match operations based on length, curvature, and separation by area between strings are then used to match candidates with the reference string. String projection, a point-by-point allocation respecting order along each string, is used if a matched string is identified. In that case, a source string may be replaced by a reference string. The use of string projection in map transformation offers several advantages. It establishes geometric consistency with the string on reference map, and hence has

detail features. It also guarantees topological consistency in a neighborhood of the matched string features. Although explicit one-to-one mapping inside a string may be temporarily lost, the bijection may be recovered by computing relative offsets inside the strings.

5) Map Transformation

Map transformation is conducted on map objects that do not have matched pairs in each hierarchical map conflation level and those transformed map objects are used as the inputs for the next level conflation. In the traditional piecewise rubber-sheet linear map transformation (White, 1985; Saalfeld, 1985, 1987, 1993; Gillman, 1985), the Delaunay Triangulation(DT), which has many unique properties, is used to provide a piecewise subdivision into triangles. Then, map objects in a triangle are transformed to the corresponding triangle. However, in the hierarchical map conflation, there is a constraint, that is, a hierarchical boundary. In order to satisfy a boundary problem and provide in more complete solutions, Weighted Delaunay Triangulation (WDT) and Linear feature based transformation (LineMorp) was also implemented. WDT was shown the best resust among three transformation methods(Figure 11).

The RMSEx, RMSEy, mean distance, range, and Standard Deviation(SD) between the selected matched pairs are used as the measure of comparison. In Figure 11, Original represents the distance differences(feet) before transformation. DT, WDT, and LineMorp represents the differnce after transformation. After transformation by DT, RMSEx, RMSEy, mean distance, range, and SD are dropped by the half of the original. In hydrography areas, the mean distance after transformation is decreased in the order of DT (213.00), WDT (188.06), and LineMorp (172.49). It is expected that WDT produces better results than DT because it uses more control points selected by the second order Voronoi diagram. The others such as RMSEx , RMSEy, range, and SD are also decreased in WDT than in DT. The more results of this proposed conflation method and implementation issues are fully provided in (Kang, 2002).

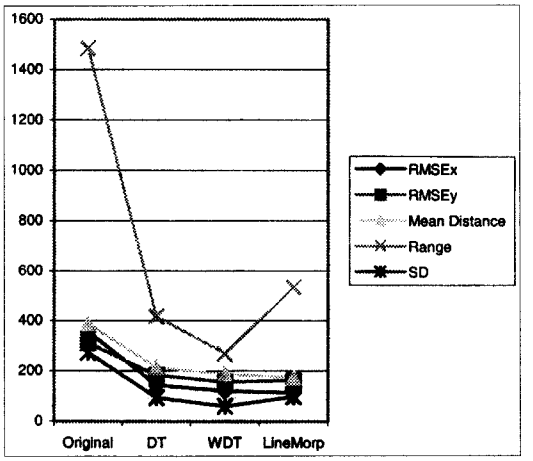


Figure 11. Comparison Before and After Transformation for Hydrography Areas

4. Conclusion

Traditional map conflation employed rubber-sheeting based on point features. In that method, the vetices of each triangle corresponds to those of conterpart triangle and the transformation of inside object is made based on those matched vertceis. The limit of this model is that it cannot correct topology differences and cannot add more linear features to represent true topology and geometry. In this paper a new hierarchically bounded map conflation incorporated with physical, logical, and mathematical boundary to reduce the complexity and computational load is presented. It guarantees producing a conflated result that is geometrically and topologically consistent with reference maps Two map models are provided for map conflation. The first model is based on the cell model of a map in which a map is a cell complex consisting of 0-cells, 1-cells, and 2-cells. The other model is based on enforcing that a homeomorphism condition is maintained among all maps. This theory helps to localize matching feature in map

conflation by assuming one-to-one relation is existed everywhere. Matching operations in map conflation are defined in two regions, that is, homeomorphic and non-homeomorphic cases. If subregions are homeomorphic, conventional matching strategy is applied. And if they are not homeomorphic, topological surgery may be implemented. *0-cell conflation test* is proposed to correct topologic difference if corresponding features are not homeomorphic.

Map conflation principles to conflate census maps consist of attribute embedding, finding meaningful nodes, cartographic 0-cell match, cartographic 1-cell match, and map transformation. An attribute-embedding component was developed to assign a spatial key to each line segment, which links to source database. A meaningful node is defined not according to geometry properties, but according to attribute properties. The categories to distinguish meaningful nodes are (1) same type and (2) same junction. Cartographic 0-cell matching uses a *0-cell conflation test* to see if the neighborhood of the 0-cell has a legitimate topology, in other words, if the two subregions are homeomorphic. Geometry, topology, and attribute match tests are performed. Using two matched 0-cell pairs and its correspondent 1-cell in the reference map, cartographic 1-cell match is implemented. Finally, map transformations are applied to transform the features inside the higher boundaries. Three different map transformation methods are implemented to study their performance in hierarchically bounded map conflation. The study shows that if an area has physically changed a lot since a real map was made, or if a real map contains heavy map generalization, then the three methods reduce the transformation distance by the half. Moreover, if an area has not changed much, or if the map contains little map generalization, then even a simple translation greatly reduce the distance. The results of map transformation may be used as inputs for the next level of conflation. Therefore, it will be easier to find matched features in the next level because map features will be more closely moved to the ground truths by transformation.

As a conclusion, this study formulates a conceptual framework and mathematical model of map conflation, develops a new match method, and provides map conflation as a tool of transformation and integration of spatial data in GIS. However, this research only examined census map conflation. There are other spatial datasets in computer mapping and GIS that need a sound conflation theory. For example, they include map revision or update(.e.g., old orthophotos replaced by new ones), maintaining map consistency of overlays, and spatial data integration. Further investigations are needed to extend conflation theory to these fields.

References

Anson, R.W., 1993, *Basic Cartography*, 2., International Cartographic Association.

Brassel, K. E. and Weibel, R., 1988, A review and conceptual framework of automated map generalization, *Int. J. Geographic Information Systems*, 2(3), 229-244.

Broome, F. R. and Meixler, D. B., 1990, The TIGER data base structure, *Cartographic And Geographic Information Systems*, 17(1), 39-47.

Buttenfield, B. and McMaster, R., 1991, *Map Generalization: Making Rules for Knowledge Representaion*, Longman, London.

Car, A., 1994a, Modelling a Hierarchy of Space Applied to Large Road Network, *IGIS'94: Geographic Information Systems, International Workshop on Advanced Research in GIS*, Monte Verita, Ascona, Switzerland.

_____, 1994b, General Principles of Hierarchical Spatial Reasoning, *The Case of Wayfinding, Proceeding of Sixth Int. Symposium on Spatial Data Handling*, 646-664.

_____, 1996, *General Principles of Hierarchical Spatial Reasoning*, Ph. D. Dissertation, Institute for

Geoinformation, Technical University of Vienna.

Callahan, G. and Broome, F., Year Unknown, *The Joint Development of a National 1:100,000-Scale Digital Cartographic Data Base.*

Census Bureau, 2000, TIGER/Line Files 2000 Technical Documentation, http://www.census.gov

Corbett, J. 1979, Topological principles in cartography, *Technical Paper* 48, U.S. Bureau of the Census.

Cuenin, P. R., 1972, *Cartographie Generale*, Eyrolle.

Douglas, D. H. and Peucker, T. K., 1973, Algorithms for the reduction of the number of points required to represent a digitized line or its character, *The Canadian Cartographer*, 10(2), 112-123.

Gross, J. and Yellen, J., 1998, *Graph Theory and Its Application*, CRC Press.

Hangouet, J. F., 1995, Computation of the hausdorff distance between plane vector polylines, *AutoCarto*, 12, 1-10.

Hake, G., 1975, *Kartographie*, Sammlung Goschen Band.

Kang, H. S., 2002, *Analytical Conflation of Spatial Data from Municipal and Federal Government Agencies*, Ph. D. Dissertation, The Ohio State University.

Katoh, N., Ibaraki, T., and Mine, H., 1978, An $O(Kn^2)$ Algorithms for K shortest simple paths in an undirected graph with nonnegative arc length, *The Transactions of The IECE of Japan*, J61-A(12), 1199-1206.

Keates, J. S., 1989, *Cartographic Design and Produnction*, second edition, Longman, Scientific & Technical.

Marx, R. W., 1984, *Developing an Integrated Cartographic/Geographic Data Base for the United States Bureau of The Census*, Bureau of the Census.

McMaster, R. B., 1992, *Generalization in Digital Cartography*, Association of American Geographers.

Mohar, B. and Thomassen, C., 2001, *Graphs On Surfaces*, Johns Hopkins University Press.

Ramirez, R., 1993, Development of a Cartographic Language, *Lecture Notes in Computer Science*, 716, Springer-Verlag, 92-112.

Robinson, A. H., Morrison, J. L., Muehrcke, Phillip C., Kimerling A. Jon, and Guptill, Stephen C., 1995, *Elements of Cartography, Sixth Edition*, John Wiley & Sons.

Saalfeld, A., 1986, Shape representation for linear features in automated cartography, *Technical Papers of the 1986 ACSM-ASPRS Annual Convention*, 1, 143-152.

_____, 1993, *Conflation : Automated Map Compilation*, Dissertation, University of Maryland.

_____, 1998, *Topologically Consistent Line Simplification with the Douglas-Peucker Algorithm*, the presented paper from Department of Civil and Environmental Engineering and Geodetic Science, The Ohio State University.

Thompson, M. M., 1988, *Maps for American*, Third Edition, U.S.G.S.

Timpf, S., 1998, *Hierarchical Structures in Map Series*, Ph. D. Dissertation, Institute for Geoinformation, Technical University of Vienna.

Topfer, F. and Pillewizer, W., 1966, The principles of selection, a means of cartographic generalization, *Cartographic Journal*, 3(1), 10-16.

Received April 21, 2004
Accepted December 17, 2004

Correspondence : Hoseok Kang, Industry Expert Center, Samsung SDS(hoseok.kang@samsung. com, phone: 011-9258-3237, fax: 02-6484-0736)

교 신 : 강호석, 463-810 경기도 성남시 분당구 구미동 159-9 삼성SDS IE센터(이메일 : hoseok.kang@samsung.com, 전화: 011-9258-3237 팩스: 02-6484-0736)