POLYGON DIGITIZING METHODS AND EXPERIENCES

J.Z. Yan
Statistics Canada

ABSTRACT

Closed regions or polygons serve as the basic unit for much of computer mapping and spatial analysis. Hence the digitization of such regions is an important operation. Several methods for the digitization and creation of area data are examined: variations of polygon, line, and segment encoding. A new method called polygon construction which combines some of the advantages of polygon and line approaches is described. For each alternative, comparative analysis is given, as well as empirical results based on experience with the AUTOMAP and GUESS on-line input systems at Statistics Canada. The effectiveness of the man/machine interface in these digitizing systems is examined.

MÉTHODES ET EXPÉRIENCES DE DIGITALISATION DE POLYGONES

RÉSUMÉ

Les régions dites fermées ou les polygones servent d'unité de base à la plupart des analyses spatiales et de cartographie automatisée. La digitalisation de telles régions est donc une opération des plus importantes. Plusieurs méthodes de digitalisation et de création de ces régions sont étudiées ici: codage de variations de polygones, de lignes et de segments. On décrit aussi une nouvelle méthode, appelée construction de polygones, qui combine certains avantages de l'approche de lignes et polygones. Pour chacune des alternatives, une analyse comparative est fournie, de même que des résultats empiriques tirés des expériences effectuées à Statistique Canada avec les systèmes d'entrée directe AUTOMAP et GUESS. On a enfin examiné l'efficacité de l'interface homme/machine dans ces systèmes de digitalisation.

POLYGON DIGITIZING METHODS AND EXPERIENCES

1. Introduction

The manipulation of two-dimensional closed figures or areas is a major concern in the fields of computer graphics, spatial analysis, and geographic information processing. Of the various representational formats for areas, described by Dueker (1972) and others (Deecker, 1974), the most frequently used description is the polygon or boundary outline.

Canadian Census operations (within Statistics Canada) have had continuing strong requirements for the digitization of polygons defining geographic areas since the inception of the geocoding or GRDSR system in 1968. In excess of 18,000 polygons defining user-specified areas have been digitized and used as the basis for retrieval of census micro-data. Another 10,000 polygons defining standard statistical areas have been digitized and processed by the CGMF system for the 1976 Census to serve the role of quality assurance of the Census geographic hierarchy. In addition, several thousand polygons have been digitized for the production of thematic maps, by use of SYMAP and GIMMS, that appeared in 1971 and 1976 Census publications.

Given the continuing demand for area digitization, the drop in costs of system technology, and the expected benefits from on-line digitization, hardware was purchased in 1977 to enable on-line input and edit of spatial data. In the past two years, several experiments have been initiated to evaluate on-line encoding and two systems, AUTOMAP and GUESS, have been put into production.

This paper reviews some of the more common methods for digitization and creation of area data in polygon form. The methods are compared based on practical experience with both off-line and on-line approaches at Statistics Canada.

2. Geoprocessing Systems at Statistics Canada

The GRDSR system (Statistics Canada, 1972) provides census information by user specified areas in Canada's larger urban centres. External users outline their areas of interest on maps. These areas are then encoded and entered in a query area library. The definitions are then passed against a micro-area centroid file and census data is extracted for all centroids falling within each query area.

The CGMF system maintains in database form the official census codes, names, hierarchy and boundaries of the standard geo-statistical areas (provinces, counties, census tracts, etc.) for Canada. One large component of the system involves the digitization and editing of polygon boundaries.

The GIMMS system (Waugh and Taylor, 1976) is a user-oriented interactive or batch system for the production of thematic maps utilizing a plotter as an output device. GIMMS has facilities for data input, redundancy checking, storage, retrieval, manipulation and display. GIMMS is now the principal production system for automated thematic mapping at Statistics Canada.

The AUTOMAP system (Systemhouse, 1978) was acquired for the public sector by Environment Canada and Statistics Canada in 1977. It is used for on-line input, edit and output of point, line, and textual data. It allows the user to create geographic base files, edit these files and output them as hard copy plots or in a digital format for further processing. Since 1977, major enhancements, implemented by Systemhouse under contract to Statistics Canada, have been directed towards the input of areal data either by segments (with automatic conversion to polygon outlines) or by complete polygon boundaries with special macro-level commands to reduce the likelihood of improperly duplicating segment boundaries for adjacent polygons.

A typical AUTOMAP work station is composed of one digitizer, one Teletec CRT terminal for command input and one Tektronix storage graphics CRT for data display. A joystick is connected to the Tektronix display. The system is driven by commands typed in by the operator, with over sixty commands currently available.

The Geographically Unique Encoding Sub System, (GUESS) (Deecker, 1978) was developed within Statistics Canada drawing on the interactive technology of AUTOMAP and the segment-oriented input methodology of GIMMS. GUESS was developed for one specific application: the input of polygon data as segments for GIMMS. The principal design objective was to improve the man-machine interface during on-line digitization based on experiences with AUTOMAP. One particular goal was to reduce the requirement for textual command input and input from more than one device.

Two large GRADICON digitizing tables are used for off-line or on-line digitization. The on-line input systems, AUTOMAP and GUESS are installed on an HP-1000 21 MXE with 64K bytes of memory. Hardware is available to operate two digitizing stations simultaneously. A remote job entry facility is utilized for the direct transfer of data between the mini-computer and the mainframe.

3. Polygon Digitization Methods

In a network or mosaic of non-overlapping polygons there is a 3-level hierarchy of point, line and area elements. Given three levels, and data structure requirements for explicit or implicit topological relationship among elements at any of these levels, a myriad of encoding schemes have been developed. Dueker (1972), for example, has listed ten alternative methods for encoding boundary data. In this paper, we limit our consideration to digitizing schemes which yield, ultimately, explicit polygon boundary outlines and/or explicit topological relationships between adjacent polygons, as these are essential requirements for both the shaded thematic mapping capacity

and the point-in-polygon retrieval facility within Census.

Since the terminology employed in the literature is not completely standardized, we shall define the terms as they will be used in this paper.

Point: a discrete location represented by a single x-y coordinate pair.

Vector: a straight line joining two points.

Line: a vector or a series of vectors joined end-to-end.

Polygon: a closed region bounded by straight lines that may be constructed from points, vectors, or lines.

When a network of non-overlapping polygons is considered, the terms below are required.

Chain: the line formed by the inclusive string of points along the common boundary between 2 adjacent polygons. For example the chain between polygons CD4701 & CD4702 in Figure 1 is (C,H,I, B). Note that polygon chains are quite distinct from Freeman chains of unit length.

Polygon junction point or node: the end point of a chain. It is normally the point where 3 adjacent polygons in the network touch.

These spatial data elements have been defined in terms of their individual x-y coordinate values only. However, topological information defining relationships between individual elements is often useful e.g. in error-checking or building structures at higher-levels. A segment as defined below requires both metric and topological information.

Segment: a chain with associated topological information describing the areas on either side of the chain.

3.1 Direct Polygon Digitization

Method: This is the most direct and perhaps the simplest method in terms of data processing. Each polygon is digitized as a sequence of points describing the perimeter of the area. Areas are encoded individually with no regard for adjacencies or the structure of constituent lines or chains.

Notes: This corresponds to method #1 in Dueker's taxonomy. Each internal chain is digitized twice and must be adjusted or averaged to eliminate gaps and overlaps or sliver lines. If this adjustment is to achieve a high success ratio, the location of points to be digitized must normally be marked on the map prior to encoding to ensure that successive digitizations of each chain will match within a specified tolerance.

Use: This method has been employed for off-line digitization of more than 25,000 polygons over the course of seven years with the GRDSR and CGMF systems. Direct polygon digitization was the method first developed for on-line area digitization with the AUTOMAP system at Statistics Canada. The MAP/MODEL (Arms, 1970) and GIDS (Yan, 1973) systems also employ this methodology. Polygon digitization is most frequently used where there is little or no requirement for segment-oriented manipulations, as generation of the constituent chains requires major processing.

3.2 Polygon Construction

Method: Polygon Construction, developed at Statistics Canada, is a variant of the polygon digitization method which removes the requirement for double digitization of internal chains. Polygons are "constructed" one at a time from a newly digitized chain and portions of existing polygons. Only polygon junction points need be digitized more than once. Thus, the possibility of slivers is greatly reduced.

Example: Consider Figure 1. The first polygon (CD4701) would be digitized using direct digitization. Encoding the second polygon (CD4702) involves digitization from point D through E to node B, duplication of the chain BC from the previous polygon, and finally closure of the polygon. The third polygon would be constructed through the digitization of a new chain (FE) and the concatenation of portions of the two other polygons.

Notes: With this method, internally consistent polygons can be produced. (no internal gaps or overlaps). Furthermore, less digitizing is required than with the direct digitization method, particularly when the number of points per chain is more than five. However, the boundary information is not maintained at the chain level. Thus, if a common chain between two polygons is to be edited, the same edit will have to be made separately for each polygon.

Use: The polygon construction method has been employed successfully at Statistics Canada (by use of AUTOMAP) to generate boundary data for the GRDSR and CGMF systems. The method has also been used to construct new polygon files following shoreline from existing polygons and shoreline features digitized separately.

3.3 Chain Digitization: Human-Assisted Polygon Generation

Method: The individual polygon chains on a map are encoded one at a time with little or no regard for the complete boundary definition of any polygon. Only polygon junction points are digitized more than once. Computer processing, with human assistance is then used to produce polygon boundary data files - usually without topological structures being generated.

Example: The operator may digitize the six chains contained in Figure 2. Then by means of a joystick he will identify, for example, the three chains that define the polygon CD4703. Straight-forward computer processing can then produce the polygon outline. This method,

implemented under AUTOMAP, is known at Statistics Canada as "polygon composition" because polygons are defined as composites of a set of digitized chains and a text feature. This and the previous two methods are described elsewhere in more detail (Yan, 1978).

3.4 Line or Chain Encoding – Automatic Polygon Generation

Method: Chains are digitized entity by entity in a manner identical to the previous method. The task of creating closed polygon boundary data from the input set of chains is handled by large-scale sophisticated software. A secondary input file containing polygon labels and an associated point known to be internal to the polygon boundary is usually supplied.

Use: This method is very efficient at encoding time, but requires complex software. For this reason it is not yet implemented at Statistics Canada. This method is used to generate polygons from scanned data by the CGIS system (Tomlinson, 1976). BNDRYNET (Douglas, 1973) is another system which generates polygons automatically from line input.

3.5 Segment Encoding

Method: The method is identical to chain digitization with the exception that for each chain two additional information elements are added: the codes or names of the polygons on the left and right sides of the chain.

Use: This is the methodology advocated by the Segment Oriented Referencing Systems Association. Segments are now the basic data structure for many popular geographic information systems. Segment encoded data at Statistics Canada is currently digitized using GUESS. In excess of 3000 polygons have been digitized by this method to date. Polygon digitization is a 2-stage process. First polygon names are entered, and associated with digitized polygon centroids. Then segments are digitized with the first two points being the centroids of the polygons on the left and right of the segment. The corresponding labels are automatically retrieved and associated with the segment.

4. Applications and Observations

Previous to 1978, production digitization of areas at Census was performed off-line using almost exclusively the direct polygon digitization method. This method permitted relatively simple clerical input procedures and file structures, both important factors in a production environment.

With the inception of on-line digitizing technology, a rapid progression was possible from the direct digitization method to the other methods mentioned in the previous section. Operators requested methods which would both reduce the number of digitizing operations and increase the consistency and quality of the data being entered. The pursuit of an optimal digitizing method for the various Census applications was the motivating factor for this work.

A detailed comparative analysis of the methods focussing on the number of digitizing operations, the types of errors, the post-processing required to produce polygon boundaries, and the ease of various graphic manipulations is the subject of another paper (Yan, Deecker, 1979). In this section the experiences to date with the various methods and systems are summarized: the comparative times, subjective analysis based on the comments of digitizing operators, and an analysis of the quality of man/machine interaction with the various digitizing systems.

4.1 Timing Analysis

Initial tests involved digitizing the census division boundaries of Saskatchewan (Figure 3) and Alberta (Figure 4), each from a single mapsheet. Timings based on the work of three different operators are given in Table 1.

Polygon construction was faster than direct polygon digitization by 15%, where the average number of points per segment was small, and by 63% on the more complex Alberta map. The time required to digitize the chains was less than that required by construction. However, the excess time required to aid in the composition of polygons, or to set-up the segment labels in GUESS, resulted in more time being used overall to input using these two methods.

A second test involving the processing of a typical Geocoding user request by three different operators using the various polygon digitization methodologies, generally confirmed the initial findings. Construction offers definite advantages in both time and quality over direct digitization.

From our experience with both off-line and on-line digitization, we report that the actual digitizing (using the direct method) may take longer on-line but overall savings in operator time are expected because of improved quality control. Definite savings in elapsed time also occurred with the transition from off-line to on-line work.

4.2 Subjective Analysis

During the previous experiments, the operators were asked to record their impressions. From their comments the following summary has been prepared.

Clerks familiar with the off-line method were pleased to be able to see a display of their work during digitization and to catch major errors immediately. However, the frequency of the system being "down" increased with the number of pieces of equipment required.

Operators preferred construction to direct digitization because of the obvious improvement in quality. In fact, operators found direct digitization unsatisfying because of the many sliver lines displayed. With construction, immediate quality assurance of the digitized product is possible. Also there is no need for post-processing to generate polygon outlines.

Operators were pleased with chain digitizing once AUTOMAP procedures were set up to provide the appearance of digitizing polygons. However, the process of manually composing polygons from constituent chains was relatively difficult for medium data volumes. Recently AUTOMAP has been extended to allow features to be automatically included in polygon composites as they are digitized. Thus, improvements in digitizing times and operator satisfaction are expected for the polygon composition method. The strongest advantage of chain and segment digitization becomes apparent when large numbers of polygon boundaries change and must be updated.

Operators found segment encoding under GUESS quick, convenient, and relatively error-free. On-line digitization of segments using GUESS was much faster than off-line digitization of segments using GIMMS because it is faster to enter a label by digitizing the corresponding centroid than to type it in.

4.3 Man/Machine Interface

One concern mentioned repeatedly by the digitizer operators using AUTOMAP was the excessive amount of textual command input required. Comments such as "I'm no typist" were frequently stated during production tests. Operating in off-line mode with output in the form of punched cards no typing ability is required. Using AUTOMAP, however, each operation is initiated by typing a command keyword and related parameters. The operator must alternate between the command keyword and the digitizer cursor or joystick. In fact, at one point production supervisors were using two clerks, one to operate the digitizer and one to type in the alphanumeric commands. Newman & Sproull (1973) note that the key to effective use of interactive systems is to reduce the input, in the main, to one device. It is apparent that this particular advice did not completely filter through to the software houses that plan interactive "million-dollar" systems for production use.

The first improvement in the man/machine interface was to set up macros or pre-arranged procedures of AUTOMAP commands for standard digitizing sequences. An increase in speed of digitizing by a factor ranging from 1.85 to 3.36 was reported (Renaud, 1978) as a direct result. After the introduction of macros the marginal benefit of using two operators decreased markedly. However, operators soon became accustomed to macros and then experienced some difficulty in using the system for cases where the standard macros did not apply. Enhancements to the basic macro facility to permit looping, and immediate exit (if something goes wrong), have further improved the man/machine interface under AUTOMAP. The success of macros rests in the fact that the operator can initiate a specific digitizing sequence and then digitize without constant command prompting. However, the operator still must frequently return to the Teletec terminal to type in macro commands.

The man/machine interface for on-line digitization could be further improved through use of function buttons on the digitizer cursor, a greater use of audio responses, and a repositioning of the graphics screen.

GUESS, programmed after initial experience with AUTOMAP, had as a specific objective minimizing the interplay between the digitizer and

the alpha keyboard. An operator would use either one device or the other, depending on requirements, rather than frequently alternating between the two input devices. Set up for segment encoding, there are basically two commands for digitizer input: one for centroid input and one for segment input. Input on the Teletec CRT has been reduced, in the main, to four keys: Q, D, N, and O which are defined as follows:

Q Terminate input loop and await another command,
D Delete the segment currently being digitized if an error on input has been committed,
N Declare the segment endpoint is a new node,
O Declare the segment-endpoint is an old node.

Target nodes that are within a specified range of previously defined nodes are identified by the system as old nodes -- and their X, Y coordinate values are modified to be the same as the previously defined node. Thus gaps and overlaps of segment endpoints are minimized. Target nodes outside the range of any old node are defined as new nodes and their actual X and Y coordinates are preserved in the node file. To be able to define new nodes close-to-but-different-from existing nodes, the operators can redefine the range value.

Audio signals are used to differentiate the detection of an old node, a new node, or an error condition. With the audio responses there is much less need for the operator to check the graphic display and hence digitization is speeded.

To summarize, a major difference in the man/machine communication between GUESS and AUTOMAP is that GUESS prompts the operator whereas, even with macros, the operator must habitually prompt AUTOMAP to accept digitized input.

5. Conclusion

Several methods for the on-line digitization and creation of polygon data have been utilized and studied recently at Statistics Canada. Chain and segment digitization require fewer digitizing operations than direct polygon by polygon digitization. The new method of polygon construction was developed as a hybrid between the chain and polygon approaches. This method has the advantage that each chain is digitized once and internally consistent polygons are constructed immediately for direct input to existing polygon manipulation programs. For this reason construction has become the preferred method of the various on-line methods considered for input to the GRDSR and CGMF systems. Production staff have indicated they would seriously consider chain digitization only once the composition of polygons is made substantially easier.

While segment digitizing and use of segment data for geographic manipualtion is more economic for some applications, its use in Statistics Canada is currently restricted to thematic maps, because existing large scale data bases are polygon boundary oriented and changeover costs would be significant.

Experience has shown that on-line digitizing offers advantages in terms of improved quality control and reduced elapsed time. Both GUESS and AUTOMAP have been used effectively for the input and edit of areal data, GUESS using segment encoding, and AUTOMAP using a variety of encoding methods. Experience with AUTOMAP has led to an improved man/machine interface in the development of GUESS.

References

(1) Arms, S. "MAP/MODEL System: System Description and Users Guide", Bureau of Governmental Research and Services, University of Oregon, 1970.

(2) Deecker, G.F.P. and Penny, J.P., "On Interactive Map Storage and Retrieval", INFOR, Vol. 10, #1, Feb. 1972.

(3) Deecker, G.F.P., GUESS User Manual, Draft, Spatial Research and Development Group, Statistics Canada, 1978.

(4) Douglas, David H. (1973), "BNDRYNET", Peucker, T.K.(ed), The Interactive Map in Urban Research, Final Report After Year One, University of British Columbia.

(5) Dueker, K.J., "A Framework for Encoding Spatial Data", Geographical Analysis, Vol. 4, No. 1, 1972, pp. 98-105.

(6) Newman, W.M., Sproull, R.F., "Principles of Interactive Computer Graphics", McGraw-Hill, 1973.

(7) Renaud, M., "Efficiency Test: Boundary Encoding Sub-Activity", unpublished report, Statistics Canada, file CEN/10526-4, August, 1978.

(8) Statistics Canada, GRDSR: Facts by Small Areas, June, 1972, published by Authority of the Minister of Industry, Trade and Commerce, Ottawa.

(9) Systemhouse (Ottawa) Ltd., AUTOMAP II Input Sub-system Station Operator's Guide, Statistics Canada, 1978.

(10) Tomlinson, R.F., Calkins, H.W., Marble, D.F., "The Canadian Geographic Information System" in Computer Handling of Geographic Data. Unesco Press, 1976.

(11) Waugh, T.C., and Taylor, D.R.F., "GIMMS/An Example of an Operational System for Computer Cartography", The Canadian Cartographer, Vol. 13, No. 2., 1976, pp. 158-166.

(12) Yan, J.Z., GIDS - A Geographical Information System, M.Sc. Thesis, Department of Computer Science, University of British Columbia, Vancouver, 1973.

(13) Yan, J.Z., "Methods Currently Available for the On-line Digitization of Area Data", unpublished report, Statistics Canada, May, 1978.

(14) Yan, J.Z., and Deecker, G.F., "A Comparison of Polygon Encoding Methods", submitted for presentation at the SIGGRAPH '79 Conference, Chicago, August, 1979.

Figure 1 An Example of Polygon Construction

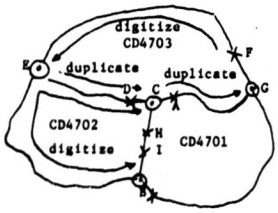


Figure 2 An Example of Polygon Composition

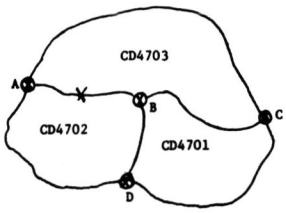
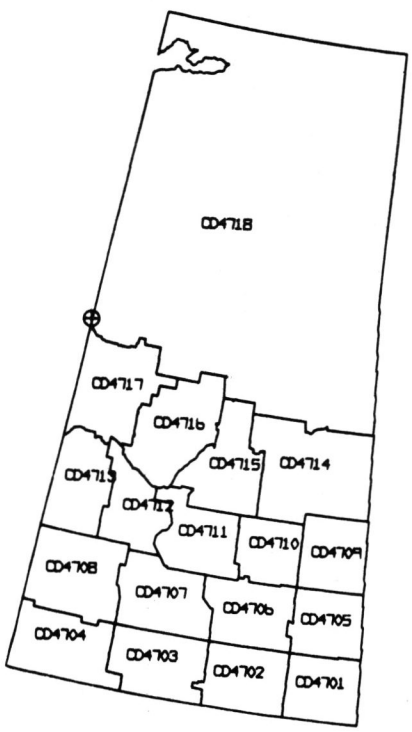
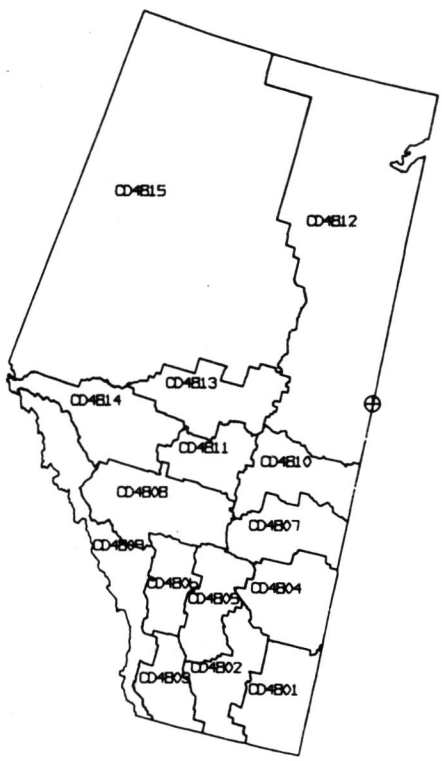


Figure 3 Census Division Boundaries for Saskatchewan



Note: This map was the subject for the first set of tests.

Figure 4 Census Division Boundaries for Alberta



Note: This map was the subject for the second set of tests.

TABLE 1

Average Digitizing Time per Map Sheet in Minutes

	DATA VOLUME					TIMINGS			
	# polygons	# inside segments (NSI)	# outside segments (NSE)	total # points (N)	average # points/ segments (V)	Direct Polygon	Polygon Construction	Polygon Composition Digitize & Compose	Segment Encoding
Saskatchewan Counties	18	35	14	362	7.4	47	41	33+12	43
Alberta Counties	15	37	10	1967	41.7	80	49	—	—