DATA GENERALIZATION AND ATTRIBUTE CODE SCHEMES

DR. MARBLE: The two topics that we are about to take up cover data generalization and attribute code schemes. Our first speaker is from the Federal Republic of Germany, Dr. Hans-Jorg Gottschalk. He comes to us with a background in geodesy and surveying. He is associated with the International Cartographic Association, and will talk to us on data generalization. Thank you.
DR. HANS-JÖRG GOTTSCHAULK: Thank you, Dr. Marble. It was about three years ago when I was here on AUTO CARTO I, and I was able to present a cartographic system which was prepared by the German Council of Research for mere research purposes. Now I am able to show you some of the work which has been done with this system. This is mainly work on generalization. Cartographic generalization is one of the most complex problems in cartography, and to find a model for cartographic generalization means to find a mathematical formula describing this very complex procedure in cartography. You could compare this to finding some kind of a cartographic world formula. There are approaches for the solution of this problem. Those of you who were in Moscow in 1976 might have read a paper given by Vasmut 1976, who tried to find a formula describing all the procedures being carried out in producing topographic maps. But nobody knows whether this formula works or not. There is no experience with this, and I think there is no experience especially in practice. Thus, the only way to solve the problem of generalization is to cut it down into pieces being so small that they can be handled. Generalization procedures are simplified, and the inter-dependencies are reduced to what is possible. The simplified procedures are mathematically formulated and computerized. This is the way computer assisted generalization has been done until now. The partial solutions and their results will be presented in this presentation. The general idea is that it is not possible to define a computer program being able to do all generalization work. It is understood that the program does the routine work of the cartographer, whereas the cartographer corrects the results of the work of the program by means of interactive editing facilities. Let me show you a slide. This is the type of data to be digitized, and this is the type of map we were able to derive from the digitized data shown before. This map has been produced by simple operations, and you will notice that, for example, there is no displacement in this type of map. You see the symbolizations are very simple. It has been adapted to what is possible in automated cartography, and also the hill shading has been done automatically, Gottschalk 1974. Now, the definition of generalization programs starts from what kind of features have to be generalized. The cartographic elements may be points and lines. Lines may be curved lines or polygons. The lines may form network structures as nodes.
and edges. In topographical maps, lines may represent surfaces or surround areas, like vegetation areas. Accumulation of points may form areas. The most common operations carried out in generalization is to smooth lines, to select lines and other cartographic elements, and to displace them after smoothing if their distance gets too narrow. There are many smoothing techniques -- for example, filtering by means of arithmetic means, splines, or dropping points of polygons (Jancaitis 1973). Smoothing can be applied to surfaces. There is an example made by Waldo Tobler in 1966, and others.

The degree of smoothing is defined by the scale to what the map is to be generalized. Although there are statistical laws giving the necessary loss of information by generalization, the last decision whether a line is properly smoothed or not is made by the cartographer. The parameters of the program are changed as long as the solution is accepted by him. This means an interactive working. If we deal with curved lines, they may be contours, roads, or lines surrounding vegetation areas. They can be smoothed, for example, by means of a gliding arithmetic mean where you use a special weight function; the degree of smoothing is due to this weight function. There have been a lot of publications about this, so I need not repeat. To smooth polygons may be a little more difficult. Usually polygons will form political borders and things like that, and political borders will coincide, for example, with other topographic items like rivers or so. And if you treat them independently you will have discrepancies in the result. So, polygons are a little bit difficult to handle. There are several methods to do this. Just drop some points out, let us say, every fourth or fifth point, or considering the length of edges between two nodes or the distance of a point from the connection between its two neighbors. If the distance is below a certain limit, the point is dropped from the polygon.

These cartographic line elements usually form network structures, which are generalized by dropping the less significant nodes of the structure. If you regard surfaces represented by contour lines, you can look at each contour line and regard it apart. You can simplify contours by smoothing, for example, each contour separately or by smoothing the surface represented by the contour lines. I will show you an example. Have a look at these contour lines. They have been derived from calculating a function using these points being marked.
there, and then systematically you can leave out the points due to the form or the shape of the contour lines, and you see the result when you successfully drop points from the surface. The first one contained about 83 points. These are the 53 points. These are 43 points, 33 points, and 13 points. You see that the main characteristics of the shape of the contour lines are kept in shape in the performance of this kind of generalization (Gottschalk 1973).

The next procedure that has to be performed is to select lines for the generalized map. You see here the drainage of a map digitized. The next slide shows the drainage transformed for the scale of one to 50,000, whereas the generalized lines were generalized for the scale of one to 200,000. You see the operation of smoothing has been carried out, and the operation of selection has been carried out. You can either do this interactively or you can just drop out lines which are shorter than a certain limit, say, two centimeters or so. So, every line which is shorter than two centimeters does not appear in a generalized map. Of course, this is a very simple model.

Next. If you come to network structures you must apply another method. The method of dropping lines shorter than a certain limit from the structure they are forming does already mean a selection. The criterion of the length, however, is not sufficient. It is possible to form a matrix of all the nodes of the structure containing the connections from one node to all the other nodes of the map. So the nodes connected, say, the intersections connected with many other intersections are more important than those being connected to only a few points of the map. The connections can be weighted according to the importance of the line -- that is say, if there is an important road intersecting with a less important road, the weight of the point is different from this. The result of all these considerations is a cardinality-matrix, the eigen values of which were already mentioned this morning, which are used for the decision whether a node of the structure is removed from the structure or not in the generalized map (Franke 1973). If you should show the next map, what you saw before was the digitizing of the scale one to 10,000. Here you see the generalized map generalized to this system that I have declared before. You see for the scale one to 25,000, one to 50,000, one to 100,000, one to 200,000, and you see that the main structure is kept whereas the less important roads or intersections
between lines are dropped from the picture. Let me show you the next slide, please. On this picture you see the main procedure which has to be applied to displacement. Applying all the procedures I have mentioned means treating a problem sequentially which is not a sequential one. As a result of these manipulations there will be intersections of lines which should not intersect, or lines will be so close together that the minimal dimensions for distances of two separate objects in a map remain below the threshold necessary for the scale of the map. At these parts of the map, the results of these simplification operations, as I have described before, have to be corrected by displacement performed either by interactive operations or by batch processing. What is to be done by displacement can be seen from this slide. You see above, the road and houses or street and houses. In changing the scale from, say, one to 25,000 to one to 50,000, you have to widen up the road. You must find an algorithm for this. One solution for this is a topological algorithm which regards this drawing as if it is on a cloth of rubber, and then you make some calculations in order to widen up the road and let the houses be apart from what they were before. You can solve this problem either by interactive solutions or by batch processing. I will show you some samples for interactive solutions. See, here some subjects which are smoothed and simplified, and you see that the houses are lying within a road. They have to be moved out so that they lie on the side of the road. This shows the display and shows the principle of the operation. You have a joystick or something like that, and by marking these two points there, you give the direction where the house has to be displaced. The result after giving a certain command is this one, so that the houses have been removed and the road is free of them (Christ 1976). Here is the problem where a creek or stream intersects with the road. The principle applied in generalization is that, generally, the creeks are kept in their situation and any other object is shifted. Here you see the operation on display, and here is the result, that has come out after this interactive operation on the display.

It turned out while performing interactive solutions that the interactive solution takes more time than we had thought before. This made it necessary to find some other solution. There is another disadvantage. If two cartographers do the displacement work at the same object at the same time or if one cartographer does the
same work twice, the result will not be identical. This made it seem useful to reduce the interactive solutions in favor of batch solutions, letting a program make the decisions, and process the whole map automatically, the result being at least uniform, is not better than the interactive solution. There are two ways for the solution of the displacement problem in batch programs. They depend on the shape of the data, either raster data or lineal data. Here you see some samples of displacement problems. You see on the left-hand side the digitized features, and on the right-hand side you see the same object after the displacement procedure, which is done completely by batch program.

The next one, please. Here you see another example. You see the road, which is on the left side, which has been digitized, and then it has been widened and the houses have been shifted. Of course, the contour lines, too, have been shifted downward (Lichtner 1977). There are some other programs which work in a similar way, but the most difficult problem is to find out the part where displacement is needed. Here, the solution is comparatively simple if the data are given in raster form. Distances of lines can be detected by means of the distance transform and other known raster operations (Weber 1977). I will show you the result. The next slide, please. Here you see a problem of displacement. You see on the left-hand side a road and a creek, and the dark object has to be moved or the road has to be moved to the left side, and the circle, the object shaped like a circle on the upper part of the picture, will have to be removed for the room is needed by the road. In order to give you an impression or how it works, you see in the middle of this slide a raster which consists of vectors which are showing the amount of displacement which has to be done in each of these points. All of the features are transformed to a raster, each of the raster points containing a vector, showing whether the two objects are too narrow, too close together or not. Where it is very black there you see that the bigger part of displacement is needed. Then you shift the whole thing according to what is necessary. The result is on the right-hand side. Here you see it once more, an enlargement. Black means that the displacement has to take place. If the vectors are smaller you see that there is no displacement necessary (Christ 1978).

Next slide, please. You see another problem which deals with generalizing, here of houses in the scale of one to 5,000 to the scale of one to 25,000. You see that
the houses have to be simplified in the way shown by
the next slide. You see the main operations that have
to be done, which are marked by the red line. You must
omit some parts of it, you must enlarge some, you must
drop some edges and so on. The result of this operation
is that, for example, houses, the areas of which are
too small, are put together and others are dropped from
the picture and so on. In order to perform this kind of
generalization the houses are digitized. All these ope-
rations are performed in the blocks which consist of
houses surrounded by the streets (Hoffmeister 1978).
I come to the end now. You see here there are digitized
houses in the scale of one to 50,000, and they have to
be generalized to the scale of one to 200,000, where
the single house representation has to be transformed
to an area representation. You see, these are the digi-
tized houses, and these are the areas which are calcu-
lated from the digitized houses. To come to a conclu-
sion: During the performance of this generalization
work it turned out that the interactive part of the
computer-assisted generalization needs far more time
than has been thought of before. The results of the
known experiments on computer-assisted generalization
show that more work has to be done on programs solving
more problems than they do now and that interactive
working should be reduced as much as possible.
Thank you.
(Appplause).

Literature:

Vasmut, A.S.: Mathematical modeling of the carto-
graphic terrain features representa-
tion. ICA, Moscow 1976

Tobler, W.R.: Numerical map generalization and note
on the analysis of geographic distrib-
utions. Ann Arbor, Michigan 1966

Gotts chalk, H-J.: Automatic generalization of settle-
ments, traffic lines, contour lines.
ICA, Madrid 1974

Töpfer, F.: Kartographische Generalisierung.
Gotha, Leipzig 1974

Jancaitis, J.R., Junkins, J.L.: Mathematical techniques
for automated Cartography. U.S. ETI,
Fort Belvoir 1973

Iwanow, V.V.: Über einige Möglichkeiten der topogra-
phischen kartengestaltung durch Auto-
mation. Geodezia i Kartografia 1, 1965

Gotts chalk, H-J.: Versuche zur Definition des Informa-
tionsgehaltes... DGK B 189,


DIGITAL CARTOGRAPHY AND ATTRIBUTE CODES

DR. MARBLE: A small change in our program. Our speaker on attribute code schemes should have been Roy Mullen from the Geological Survey. Instead, Dr. Robert McEwen is going to talk to us on this topic. Bob is supervisory research cartographer with the Topographic Division USGS, and is currently leader of the Division's Digital Applications Team.

DR. ROBERT McEWEN: I appreciate the very short and to-the-point introduction. As has been noted, I am not Roy Mullen. However, he is sitting in the second row and if there are any difficult questions asked of me at the end of this presentation, I will be glad to refer them to Roy. I am with the Topographic Division and the Digital Applications Team (DAT) and would like to speak about some of the things we are doing and how that relates to data you may acquire from the USGS National Mapping Program. Most of our data distribution will be through the National Cartographic Information Center (NCIC). I do not know whether they have any digital related information in their workshop here but during the next year or so they will be incorporating digital data in their information.

I would also like to say a word about previews of coming attractions. Not to distract from this meeting where we have the privilege of speaking informally, but both Roy and myself have prepared papers for the ASP-ACSM meeting in Washington at the end of next month. There is also a digital terrain model symposium in May in St. Louis and Dr. Elassal of the DAT is presenting a paper on data management. Another paper will be presented by Brunson and Olson on digital terrain model accuracy and the equipment used by the Topographic Division to gather data.

We are organizing our data in two distinct types of files. As you might expect, being in the cartographic surveying profession, we are organizing horizontally and vertically. We have files that we call DEM's, digital elevation models, and we also have files which we call DLG, digital line graph, for the horizontal information. We are gathering this information with a number of pieces of equipment. There is not enough time here, nor is it perhaps appropriate, to go into a great deal of detail. But presently a lot of the horizontal information is being gathered with manual digitizers. The vertical data that is going into the DEM files is being primarily gathered with Gestalt Photomappers (GPM-2); we also have some other orthophoto instruments digitized. The GPM-2, in the process of making an orthophoto, creates digital elevation data on tape. We also have two M&S editing systems, Gerber and Calcomp plotters, and we are in the process of
installing additional digital plotters at this time.

Back on the DLG and DEM files. We have categorized those in three levels. The first level is basically unedited data -- I will not say "raw" data because sometimes we have taken the raw data and done a little bit of clean up to get it organized. But basically, Level I is the data as it has been gathered. The Level II data represents an editing function performed on the Level I data. Blunders, such as lines that do not quite intersect and lines that cross where they should not, have been edited to make a clean file. The third level of data represents, in the DEM case, elevation data that has been tied to planimetry -- in other words, the elevations have been tied to planimetric features such as roads, streams and so on, so that you have a consistent set of data. It is a normal function when you are making a map to make the contours cross the streams at the appropriate places and so on. It is the same conceptual idea in the Level III for the DEM.

In the DLG case, Level III represents a topological editing of the data, so that we have a perfectly consistent graphic that is topologically edited. In other words, as you walk around all of the nodes, you will have a consistent set of areas that you cross and come back to the same area. If you follow around polygons you will always have the same polygon on your left, and you will come back to the starting point. We think this is important, and it represents a commitment of the Topographic Division to service geographic information systems which, we believe, require topologically edited data.

For purely graphic purposes, many of the Level II types of data would be appropriate. I think it has been said before, that we are not convinced we want to digitize graphics just to make additional graphics. It may turn out as time goes on, that we will get a feedback from the digital files to speed up, modernize, and increase efficiency in our map-making operations. At the present time this may not be completely cost effective, but we believe there is a commitment to gather data that can be used for geographic information systems. And we believe that to do this we need DLG-3 data.

We have a set of computer programs that have been developed for processing planimetric data. The data gathering procedures and the processing result in the final output of an organized and formatted DLG-3 file. The programs are called the Unified Cartographic Line Graph Encoding System (UCLGES). (Nobody could pronounce that, but the closest we could come was "Uglies").
is presently operational on our IBM 370/155 and we made a conscious decision that we would only have that program in one place so that we could maintain it.

Documentation and information about the program has been written, and there is no reason why that cannot be shared with the community, although I do not know that it has been published in any form. The program requires the operator on the digitizing table to encode the nodes, and then encode lines and then encode the areas, for a planimetric graphic. We work one category of data at a time. The program goes through and performs various topological edits, throws out data that is not correct, produces a proof plot, and then the operator goes back and adds and deletes until he gets a final graphic that represents the map that he is digitizing. Finally, an output file is created, and it would be this file which would be available through NCIC. A file is organized into lists of nodes, lines and areas. Each of these can have attribute codes connected with them, which I will describe in a moment.

We are presently gathering data for a limited number of pilot projects and are proceeding very deliberately to test our own capability to gather data and also to test the cartographic community to see if it can really make use of the data that we gather. We are digitizing basically from 7 1/2-minute maps. One project, recently completed in the State of Idaho, involved some 212 quadrangles in which we digitized political boundaries, the public land net, and some forest boundaries. The purpose of the project is rather interesting, because it relates to a forest inventory in which the idea was to classify LANDSAT pixels multi-spectrally and then automatically drop those pixels into the right county or the right public land section or the right forest ownership, and to do this all in a computer without ever having to overlay a graphic.

Another project is being conducted with the Bureau of Land Management, the U.S. Forest Service, and the Bureau of Fish and Wildlife. It is in southeast Oregon and includes 38 quadrangles. We are digitizing boundaries, transportation, hydrography and some cultural features. We have another project with our Geologic Division within the USGS which relates to coal resources. We have digitized DLG data for the public land net and political boundaries, and have also gathered DEM data. We have several other pilot projects that are either just getting started or being considered. As those materialize we will gradually make that information available. I will not promise that all of this pilot
project data will be distributed. There is always the chance that some of it will be erroneous or for one reason or another is not complete.

Now, I should say a word about the National Mapping Program and the base categories of data that we have jurisdiction over and to which we are primarily paying attention. If you look at a 7 1/2-minute map you will see the base categories. They are very obvious, but you ought to stop and think about them for a moment. What is on a 7 1/2-minute map? There are the contours, some boundary information, hydrography, and transportation features such as railroads, roads and trails. You will see some information about surface cover, such as orchards and vineyards as well as the woodland tint. You will see hydrography both as streams and water bodies and wetlands. There is other information such as names and also information about the coordinate system - the state plane coordinate systems, the UTM, and the geodetic coordinates.

These are all proper ingredients for a 7 1/2-minute map and constitute the base categories of information we are responsible for in the National Mapping Program. It is to these that we will primarily pay attention in our digitizing. We are basically a cartographic data gathering, maintenance, management and distribution organization. We intend to carry through and follow these types of functions when we move into the digital domain.

Attribute codes have not been discussed too much here but we have given them quite a bit of thought, and have prepared a document that describes some of our codes. A small part of that will be in the paper that I am giving at the ASP-ACSM meeting. It is perhaps that third level that was described earlier in the day when civilization settles down to law and order and locks everything into some defined numbers.

We have code numbers that are basically seven digits, separated into a three-digit major code, a period, and a four-digit minor code. The major code, the first three digits, refers to the base categories of information that I previously mentioned -- the categories of the National Mapping Program. For instance, there is a code for transportation, a code for hydrography, and so on. The minor code, four digits to the right of the decimal point, has unique numbers to describe node information, line information, and area information. It is rather interesting to consider what you want to encode as a line and what sort of a number should you give to it. Let me give you a few examples.
On our maps we show divided interstate highways. There is a definite symbol for highways that have a median of a certain width. It is perfectly appropriate not to digitize both of those roads, but simply digitize down the middle of the median strip, and assign a code, "divided highway." You also run into a different consideration when the road divides and each lane goes its own separate direction. You now have to digitize each direction of traffic, and you have to label it an interstate highway with restricted access and one-way traffic. Then a problem came up: can we possibly convey in numerical attribute coding the direction of the traffic? Is there a number that says I am traveling this way or that way along the line? We decided in that particular instance there was probably no way; somebody is simply going to have to look at the graphic to find out which way the traffic goes. All we have said in the numbers is that it is restricted to one-way traffic.

There are many other cases of this. In the hydrography area we started to consider the questions that would be asked in the computer about hydrography, and we realized in the water resources area that there are many routing types of questions. Since we are digitizing streams in one file and the water bodies in another file, you can come to a situation along a stream where you do not know where to go to connect with the next stream that flows out of the water body. So we are forced to put in an artificial line across the water body which maintains connectiveness between the in-flow and the out-flow of a water body. We also realized that we have to digitize the banks of streams as well as the center lines where we have double-line streams. We also have situations where we have to digitize on either side of islands, because there are alternate routings in getting down the stream. We have done a fair amount of thinking about this, and I am not going to go on any further in this particular meeting, but I would call your attention to the document that we have prepared, and we would welcome your comments and criticism and suggestions as to how this data can be made most useful to people who have applications for it. Obviously one function of an attribute code is to search on it so you can pull certain information from the digital files.

Moving on for just a moment to the digital elevation models, which I indicated were being gathered primarily with the GPM-2. We have over a thousand of these quadrangles digitized. The GPM-2 generates more data than we feel is necessary to put into the files. We are presently going through a data thinning, organizing and filing procedure and establishing a management
system so the data can be distributed. We are starting to do this with some of our backlog and I anticipate either late this spring or early summer that some of these files will be available through NCIC. The files are generally regular arrays but there is a provision in the file structure for irregular arrays of digital elevation models. So, we have accommodated both types.

In closing, I would mention two other items; NCIC distributes the DMA digital terrain tapes from the 1:250,000-scale maps, and we also have contours for the Jewel Ridge, Virginia quadrangle digitized as an experimental data set. Thank you. (Applause.)

DR. MARBLE: Thank you, Bob. That was an interesting discussion, and I think enlightening to all of us in terms of its promise for the future. Are there any questions for these speakers?

MR. EDSON: I would like to add a footnote for the record. I would like to recognize a couple of gentlemen attending AUTO CARTO III who have been involved in much of the work Dr. Gottschalk described. These two gentlemen are from the Institute for Applied Geodesy in Frankfurt, Theodor Johannsen and Fred Christ. Would you stand up, please? You may want to contact these two gentlemen in the event also that you have questions later on. Thank you.

MR. LEN JARVIS: I am with B. C. Hydro Power Authority. I would just like to hand Dr. Gottschalk this little sketch. It sort of illustrates the problem which I am proposing. The sketch illustrates a fairly typical utility mapping problem where you are digitizing a facility's map which shows, let us say, gas mains and electric underground in the same road allowance, in the same street at a digitization scale of, say, one inch equals 100 feet. The actual separation between a line representing the gas main and the line representing the electric underground is really quite small. These two facilities may be in real terms only a foot apart. Now, the difficulty is that if it has been digitized at that large scale, and then the wish is to display the information on a smaller scale, let us say, one inch equals 200 feet, then on a conventional CRT, you run into the problem of lines actually merging, the distinction between them being lost. Now, if I have understood Dr. Gottschalk's generalization discussion properly in terms of the data shifting, he is able to put attributes to the information such that when the road allowance will be re-scaled from one inch in 100 feet to one inch in 200 feet, he can in fact maintain
the same visual separation between the tightly spaced gas main and electric underground cable so that the separation will still show properly at the smaller scale. I hope Dr. Gottschalk understands the inquiry, and maybe he can comment.

DR. GOTTSCHALK: Yes, I think I understand it. The problem is to which of these two lines you want to keep in place.

MR. JARVIS: I showed a solid and a dotted line. Let us say for the sake of argument that the solid line should always show to scale at the correct offset from the road allowance, and, therefore, the dotted line should always be, say, at a fixed display scale from the solid one.

DR. GOTTSCHALK: It depends on what we call the minimal dimensions, the minimum graphic dimensions which bring the two lines apart. If you refer to graphical minimal dimensions, you must draw the two lines 0.3 millimeters apart. What you must do before is decide which of the two lines you will keep in its place and which of them will have to be displaced.

MR. JARVIS: So, you are saying that given I make the choice initially, then I can in fact solve that problem.

DR. GOTTSCHALK: Actually, we did not think about this problem as far as some kind of cadastral work, I think, less than topographical mapping.

MR. JARVIS: Yes, it is definitely not a topographic problem.

DR. GOTTSCHALK: It is up to your definition of how to do it. I suppose that you can use this type of operation for this.

MR. JARVIS: That is very interesting. I won't take any more time here, then, but I should for my own edification follow that up with you at some time. Thank you very much.

DR. GOTTSCHALK: Perhaps you might want to discuss that with Dr. Johannsen and Christ.

DR. MARBLE: Do we have other questions?

DR. KURT BRASSEL: Kurt Brassel, SUNY, at Buffalo. I have a question for Dr. Gottschalk. You have shown a topographic map, a generalized topographic map, which includes several map elements. How did you correlate the various generalizations of the various map elements?
I understand right that this may have been done by a raster structure?

DR. GOTTSCHALK: My idea concerning the first map was, if there are objects being correlated in the digitized map, and if you process them with the same operation -- let us just say if you take a gliding arithmetic mean with the same parameters to all of the features, you can easily calculate it mathematically. If there was a correlation before, there must be a correlation after this application of the gliding arithmetic mean. It is just transformed. But there is still a correlation. This is proved when we, for example, smoothed the contour lines and the drainage using the same parameters in this gliding arithmetic mean formula. It proved that the drainage was still running in the valleys. Perhaps when you have been in Frankfort you have seen this large map there, and you could see that they really remain inside, the range that was expected.

DR. MARBLE: We are finished for the day. Thank you.