Current Research (1999) Newfoundland Department of Mines and Energy
Geological Survey, Report 99-1, pages 71-94

GeoSymbol: A DATABASE SYSTEM FOR MANAGING STRUCTURAL SYMBOLS IN A GEOGRAPHIC INFORMATION SYSTEM

S.P. Colman-Sadd
Regional Geology Section

ABSTRACT

Structural point data are an integral part of most geological maps. They indicate the strike/trend and dip/plunge of individual structures measured in outcrop. Most geologists collect more structural data for a map than can be shown at the intended publication scale, and this cartographic problem is exacerbated if the map is plotted at a reduced scale. Increasingly, structural data are entered directly into computer files where they are much more readily available for publication than in the past. There is also an increasing demand for customized maps generated from original datasets by someone other than the geologist who actually made the map.

GeoSymbol *is a prototype database system for storing structural point data and for producing output text files that can then be plotted directly in a Geographic Information System (GIS). It allows structural data to be plotted legibly on a map, independent of scale, and provides a geologically rational way of fitting the data into the available plotting space. It is also independent of the coding scheme used to indicate structures, allowing data to be entered using any scheme and the output to be matched to a preferred symbol set.*

The user selects a subset of data, usually for the area of the required map, and indicates what kinds of structures should be plotted and which structures should have priority in the event of spatial conflicts. There is a choice of three plotting formats, allowing just one symbol at a station or up to four symbols. GeoSymbol *determines what stations and symbols can be plotted on a map without overlap, using the requested scale and symbol size. Where there are overlaps, it eliminates stations of least value using four geological based criteria to determine which stations are least necessary. The output file gives plotting locations for all structural and outcrop symbols on the map with positions and offsets calculated to match the specified scale. No further processing is required in the GIS.*

INTRODUCTION

An essential task during bedrock geological mapping is to measure the attitudes of structures identified in rock exposures. The measurements are shown on geological maps by a variety of symbols depicting different types of structures. The orientation of a symbol indicates the strike or trend of a structure and the dip or plunge is shown by a number. The symbol may be placed exactly on the location where the measurement was made or may be offset, in which case the actual location may be marked by an outcrop symbol and several measurements may be shown for one exposure. Structural measurements are one-dimensional point data and have specific locations, which can be defined by longitude and latitude or, in Canada, by northing and easting on the UTM grid.

Structural symbols have usually been considered a standard component of geological maps at 1:100 000 or more detailed scales. At these scales, they add significant information and, in most cases, are fairly representative for the area on the ground occupied by the symbol on the map. As scales decrease, symbols become less representative because they occupy progressively larger areas. Symbols are often left off maps at the 1:250 000 and 1: 500 000 scales and are rarely included on regional compilation maps at 1:1 000 000 or less.

The number of measurements that can be included on a map, without overlaps, is limited by the scale at which the map is plotted. Larger scales allow room for more symbols. A geologist commonly makes many more structural measurements than can be plotted at the scale of the final published map and normally selects measurements for this map that best illustrate a preferred interpretation of the geological structure. Before the adoption of computers for geological field work, the remaining measurements would usually be stored in the geologist's notebooks, where they might be available to the public (in a rather inaccessible form), if the notebooks were archived. In recent years, field data have

been entered on a routine basis into digital databases such as Fieldlog (Brodaric, undated). Consequently, structural measurements have become more accessible because digital files are more readily published and because the data are organized in a standard format that is easier to use.

Despite these improvements, archives of structural measurements remain of limited use. They are still generally viewed as secondary to what the mapping geologist deemed to be the best selection for the published map. Although the complete dataset may be available as a digital file, it is difficult to use it on maps without creating unreadable congestion. Congestion can only be relieved by expanding the scale of the map or by thinning the data. Most maps are made with a particular scale in mind for a particular purpose and so expanding the scale is not desirable. Thinning the data allows maps to be produced without compromising other requirements such as scale and types of structures to be shown. However, it is time consuming and requires specialist judgement to decide which measurements to omit and which to keep.

GeoSymbol has been designed to automate the thinning process so that the user of the data only needs to specify some general preferences and the computer deals with the problem of producing the best fit of the best data for the desired purpose. This requires that *GeoSymbol* make two kinds of decisions. The first of these is spatial and determines the fit of the symbols using geometrical calculations. The second is geological and decides which symbols and stations will be retained and which discarded using geological criteria. The criteria are intended to mimic those that the mapping geologist would use, perhaps subconsciously, when thinning the data by hand. It is clear that the computer is unlikely to be as good at thinning data as the mapping geologist, both in terms of spatial and geological decisions. However, *GeoSymbol* is based on the premise that the computer can do an adequate job and that this is better than the probable alternative of not using the archived datasets at all. The need for an automated solution is best appreciated if we consider two of the big advantages provided by computerized geological maps. First, there is the ability to customize individual copies of a map for a specific need using only data relevant to that need. Second, there is the potential to access all the data that were collected during mapping, not just what could be shown on the author's original map at the scale of publication. It is reasonable to expect that the mapping geologist will select data for the initial publication version of the map. However, it is clearly not possible for the same geologist to select data for numerous versions at different scales, with different structural emphases, and for purposes that may only appear many years in the future. Unless measurements can be thinned automatically, it will usually be impractical to customize the structural symbol layer of a map and most users will be restricted to the subset of data shown on the published version. As a result, much of the data collected during mapping and entered into digital databases will be rendered unusable and redundant for most purposes.

GeoSymbol has been designed to address the problem by meeting the following objectives:

1) Stations are thinned so that the number of symbols on the map is limited to what can be plotted without overlap of one symbol on another.
2) Thinning automatically adjusts for the scale of the map; smaller scale maps can accommodate fewer stations than larger scale maps.
3) Thinning is done using geological based criteria.
4) The thinning criteria can be individually weighted or suppressed.
5) The types of symbols to be plotted on a map can be specified.
6) An order of preference can be set for the types of symbols to be plotted, both globally and at individual stations.
7) Different plotting formats allow for plotting between one and four symbols at a station.
8) Data can be entered and stored using various symbol-coding schemes in a single database.
9) The symbol codes, and therefore the symbology of the output, can be varied without having to revise the data.

The first seven objectives allow the user to adjust the output of structural symbols for maps of different scales and different themes. They complement the objectives addressed by *GeoLegend* (Colman-Sadd *et al.*, 1997), which provides the ability to adjust the level of detail shown for map units and their accompanying legends. It is intended that *GeoSymbol* and *GeoLegend* will operate together on different layers of the same GIS.

The ability to use various symbol schemes (objectives 8 and 9) relates to an important issue addressed by both *GeoSymbol* and *GeoLegend*, that of consistency between neighbouring maps. In the case of map units, there is seldom consistency from one source map to another and dealing with this problem is a fundamental feature of *GeoLegend.* For structural symbols, the situation is usually better because individual agencies tend to standardize on one symbol scheme. Nevertheless, inconsistencies can arise, especially where maps cross the boundaries between different jurisdictions. For this reason, *GeoSymbol* is set up so that it is independent of any particular coding scheme or symbology, and can, in fact, use any number of these simultaneously.

SYSTEM OVERVIEW

SYMBOL LAYOUT

GeoSymbol is a database system that interacts with the graphical component of a GIS through imported and export-

ed text files. The output text file tells the graphical component what symbols to plot, how to orient them and where to place them. *GeoSymbol* uses the same conventions for recording structural data as Fieldlog and its specifications for symbol layout are all met by the NATMAP symbol set (Brodaric, undated, Appendix 8):

1) The "right-hand rule" is used for measurements. The azimuth of the strike for a planar structure is given as the direction in which the geologist is facing when a structure dips to the right. The azimuth of the trend of a linear structure is the down-plunge direction.
2) The azimuth defines the long direction of a symbol.
3) All structural symbols have a standard length. Symbols can be shorter than the standard but *GeoSymbol* considers these to take up the same space as full length symbols.
4) Ornamentation on the right side of a symbol should plot entirely within the circle described by the standard symbol length.
5) Ornamentation on the left side of a symbol should only extend a distance from the centre equal to 20 percent of the standard symbol length. This restriction leaves room for an outcrop symbol and need only be followed if the offset plotting format is to be used (*see* "Final Changes and Plotting Formats").
6) The number giving the dip or plunge of the structure should plot on the right side of the symbol and should plot entirely within the circle described by the standard symbol length.
7) There are no special symbols for horizontal or vertical structures. These are simply indicated by 0 or 90 adjacent to the inclined symbol.
8) The location given in a *GeoSymbol* output file is at the centre point of a symbol for both planar and linear structures.
9) Outcrop locations are indicated by an isometric symbol such as an 'x'. The diameter of this symbol is limited in relation to the standard symbol length in order to prevent overlap with other symbols (*see* "Station Geometry" under "Structure of the Database").
10) *GeoSymbol* manages the generation of outcrop symbols and the calculation of offsets internally. It expects that all symbols will be plotted exactly where stated in the output file. A symbol consists of the symbol itself and a number indicating the dip or plunge, where applicable; associated outcrop symbols are treated separately.

DATA INPUT

GeoSymbol uses two data tables for recording station locations and structural measurements, respectively. These tables are deliberately similar to those used in Fieldlog for these purposes. The 'Station IDs' table (Figure 1) contains a unique number for each station, the 'Station ID', and UTM coordinates to fix its position ('Easting', 'Northing' and 'Zone'). The 'Measurements' table (Figure 2) contains 'Station ID', which links to the 'Station IDs' table, 'Symbol code', indicating the type of structure, 'Azimuth', which gives the strike of planar structures or the trend of linear ones, and 'Dip/plunge' for the dip or plunge of a structure.

The two tables also contain several variables that are not normally present in a Fieldlog dataset. These variables are used to determine the geological criteria for filtering stations. In the 'Station IDs' table, information is required about the map unit polygon in which each station is located. 'Map ID' uniquely identifies the geological map with which the station is associated and 'Polygon ID' identifies the polygon within that map. These two variables correspond to variables in *GeoLegend* and can be imported from the *GeoLegend* layer of the GIS. At the same time, the areas of the polygons can be calculated and the results loaded into the 'Polygon area' variable.

In the 'Measurements' table, 'Preferred priority' allows the relative importance of measurements at a station to be recorded. 'Key measurement?' allows a measurement to be flagged as especially important relative to all other measurements in the database, whereas 'Suppress measurement?' prevents a measurement from being used, perhaps because it is of suspect quality, but allows it to remain in the database.

A third input table, the 'Symbols' table (Figure 3), is used to define symbol codes and allows translation from one coding scheme to another. For each value of 'Symbol code', this table stores the full name of the structure in the 'General category' and 'Specific category' fields. The table also stores the 'Symbol scheme' to which the code belongs. A fourth table contains bibliographic references for the data in the first two tables and has two subsidiary tables to store related NTS information. It is the same as the 'References' table used in *GeoLegend* and, if *GeoLegend* and *GeoSymbol* are used together in a single GIS, one 'Reference' table or even an outside bibliographic source would suffice for both. The 'Map ID' field in the 'Station IDs' table links to the 'References' table and allows retrieval of the 'Map reference' and the 'GIS ID' for a particular station in both the 'Station IDs' and 'Measurements' tables.

DATA OUTPUT

Station Selection

The process of producing a structural symbol layer for a map is completely controlled from the 'Output Selection' form (Figure 4). The first task is to select which stations are to be considered for plotting. Normally this is done in the graphical component of the GIS by outlining the area of the intended map and downloading a text file of all the 'Station IDs' contained within it. The list is loaded into *GeoSymbol*

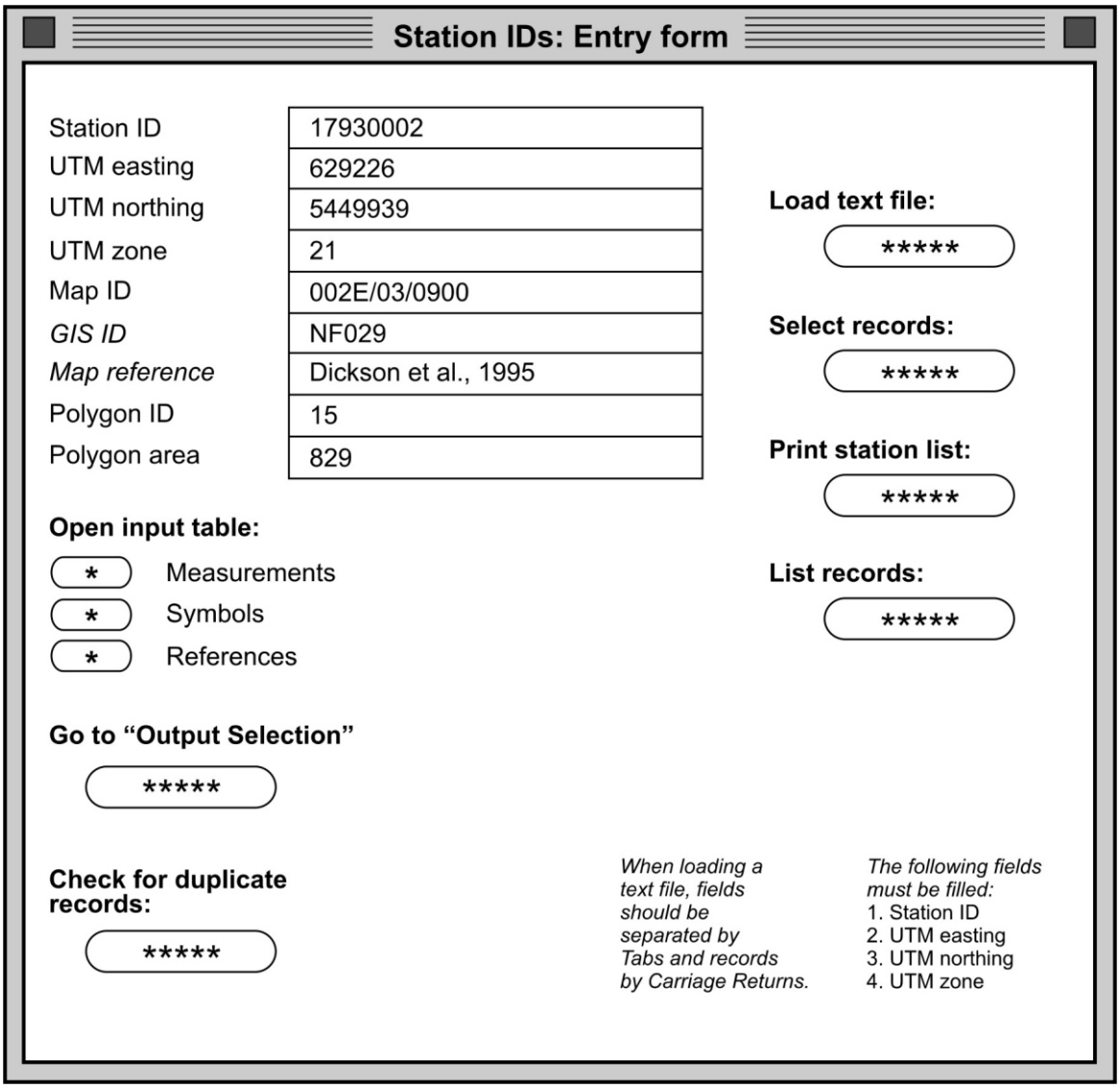

Figure 1. *Entry form for the 'Station IDs' table. Labels in Roman type indicate variables entered directly into the table. Those in italics indicate fields computed from data held in other tables.*

using the command button, 'Load station identifiers from a text file'. Alternatively, stations can be selected directly from the 'Station IDs' table using a query form, which allows areas to be defined by UTM coordinates or by specific ranges of 'Station ID'.

Spatial Parameters

The user must define certain parameters for the map that is being created. The problem of plotting structural symbols on the map without any overlap is initially a spatial one. *GeoSymbol* uses metres measured on the ground, expressed as the northings and eastings of the UTM grid, to fix the locations of stations. From this starting point, it is convenient to use the same units to do all spatial calculations required to prevent overlaps and offset structural symbols around outcrop locations. In order to do this, four basic pieces of information are needed. The first of these is the scale at which the map will be plotted. The other three define the dimensions of the symbols as they will be plotted on the map, specifically the length of a standard structural symbol, the diameter of an outcrop symbol, and the amount of clearance needed to prevent symbols from actually touching. A line that is 6 mm long on a 1:50 000 map represents a distance on the ground of 300 m. If the scale is specified at 1:50 000 and the symbol length at 6 mm, *GeoSymbol* treats symbols as if they were 300 m long when it does spatial calculations (Figure 5). If the scale is changed to 1:250 000, 6 mm symbols are treated as if they were 1500 m long. All of the offsets required to position structural symbols around outcrop locations can be calculated from the symbol dimensions and are likewise converted to metres on the ground.

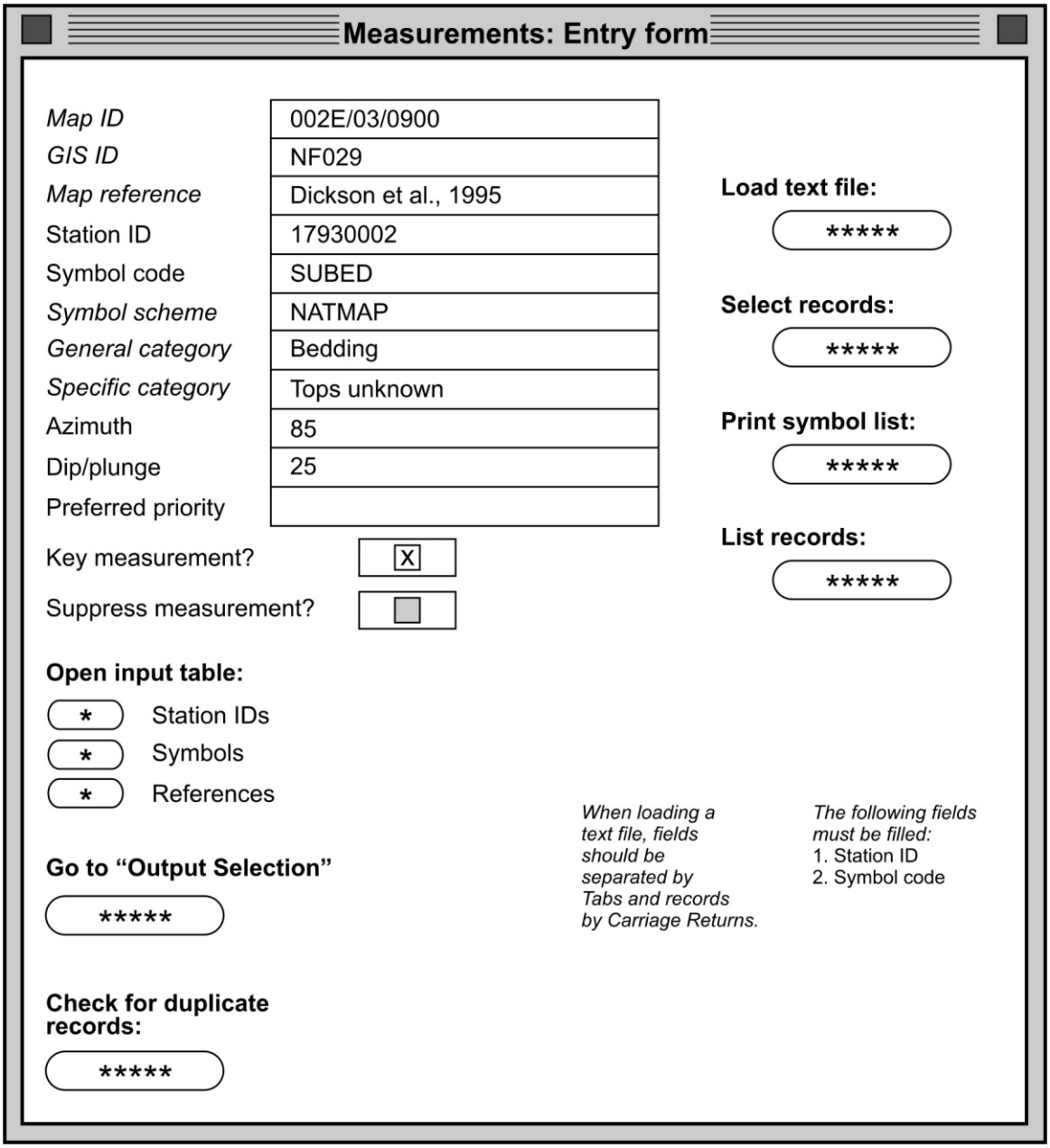

Figure 2. *Entry form for the 'Measurements' table.*

Overall Symbol Priority

A window on the 'Output Selection' form displays a list of all the structures that are represented at the chosen stations (Figure 4). An order of preference must be set for these structures so that *GeoSymbol* knows which symbols to plot and which to omit if there are too many symbols for the available space at a particular station. The structures are numbered from 1 onward, with 1 indicating the highest priority and so on. The overall symbol priority can be overridden for individual stations by making special entries in the 'Preferred priority' field in the 'Measurements' table. If the overall symbol priority is left empty for a structure, the corresponding symbols are omitted from the map, allowing thematic maps to be constructed, which only display certain types of structural information.

Weightings for Thinning Stations

Four sets of criteria are used to eliminate stations when there are too many symbols to fit on the map without overlap. The criteria can be weighted by choosing a percentage from the pop-up menus. Values range from 100%, which is full on, to 0%, which is turned off.

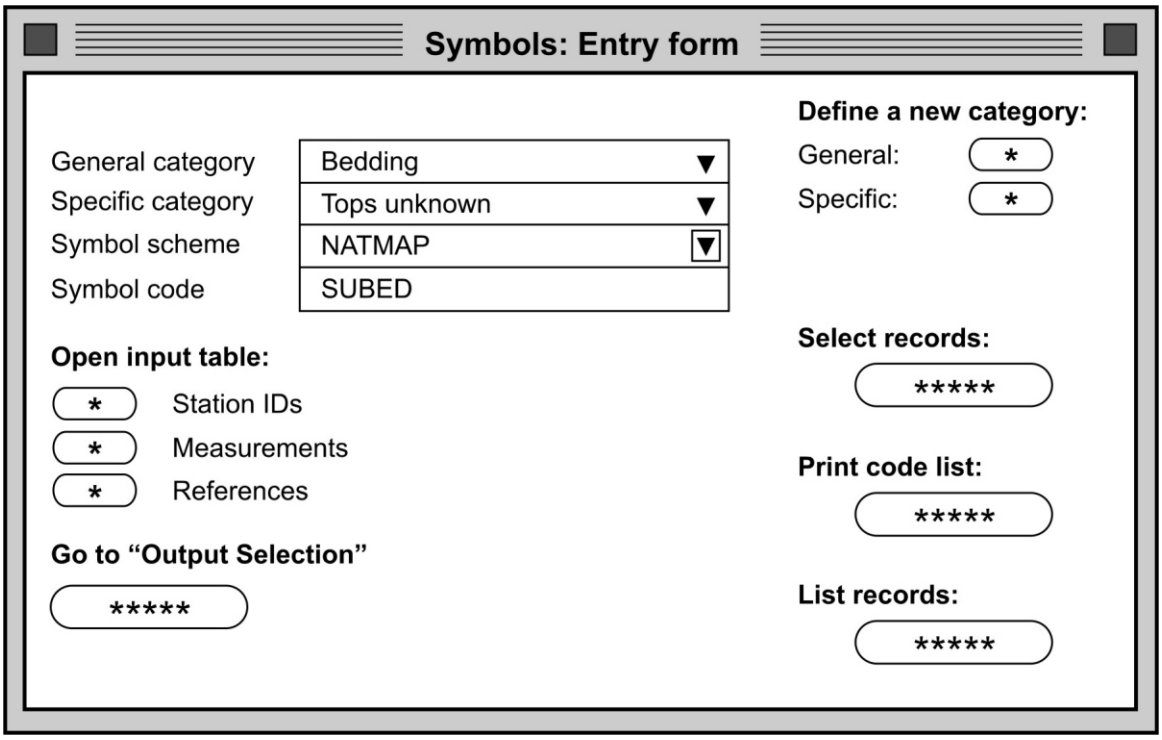

Figure 3. *Entry form for the 'Symbols' table.*

'Key measurement?' gives preference to stations for which the 'Key measurement?' variable has been flagged for a structure in the 'Measurements' table. This usually ensures that the station survives the thinning process unless neighbouring stations also have flagged measurements, in which case, thinning will be governed by the other criteria.

'Complete dataset' favours stations that have more measurements over those that have fewer, provided that the type of plot requested allows for more than one symbol at each station. Two measurements at the same location are deemed to be more valuable than two similar measurements at separate locations because they carry implied information about the relationships of the structures. For example, bedding and cleavage measured at a single location provide information on their intersection and hence on related fold axes. 'Complete dataset' also incorporates the preference for one type of structure over another, as described under "Overall Symbol Priority", favouring stations with higher priority structures.

'Representative data' aims to retain the symbols that are most representative of the various structures within each polygon. In calculating what is most representative, measurements are grouped together if they all refer to the same basic concept. For example, all structures that indicate the paleohorizontal, whether they be bedding, pillow attitudes or flow tops, would normally be compared, whereas foliations of different generations would be considered separately. The user has the option to change groupings in a subsidiary of the 'Symbols' table, if desired.

'Scarce data' is intended to protect stations that have measurements for unusual structures. If a polygon contains twenty stations with measurements of bedding and the first generation foliation, but just one station with bedding and a first generation fold axis, 'Scarce data' favours the station containing the fold axis measurement because of the scarcity of this type of data in the polygon. 'Scarce data' also gives preferred treatment to stations in polygons containing little structural information relative to their area if their symbols overlap symbols in adjacent polygons with abundant information. The calculation of 'Scarce data' involves the area of polygons and the numbers of symbols of various types in each polygon. It requires that limits be set on the area for

Figure 4. *'Output Selection' form containing all the controls needed to produce a customized symbol layer for a map at any scale. Although there are measurements of igneous layering in the selected area, they will be omitted from the map because they have not been assigned overall priority numbers. Furthermore, the symbol for "Igneous layering, tops unknown" has not been defined for the NATMAP scheme in the 'Symbols' table. The counter at the bottom of the form provides a continuous status report during processing.*

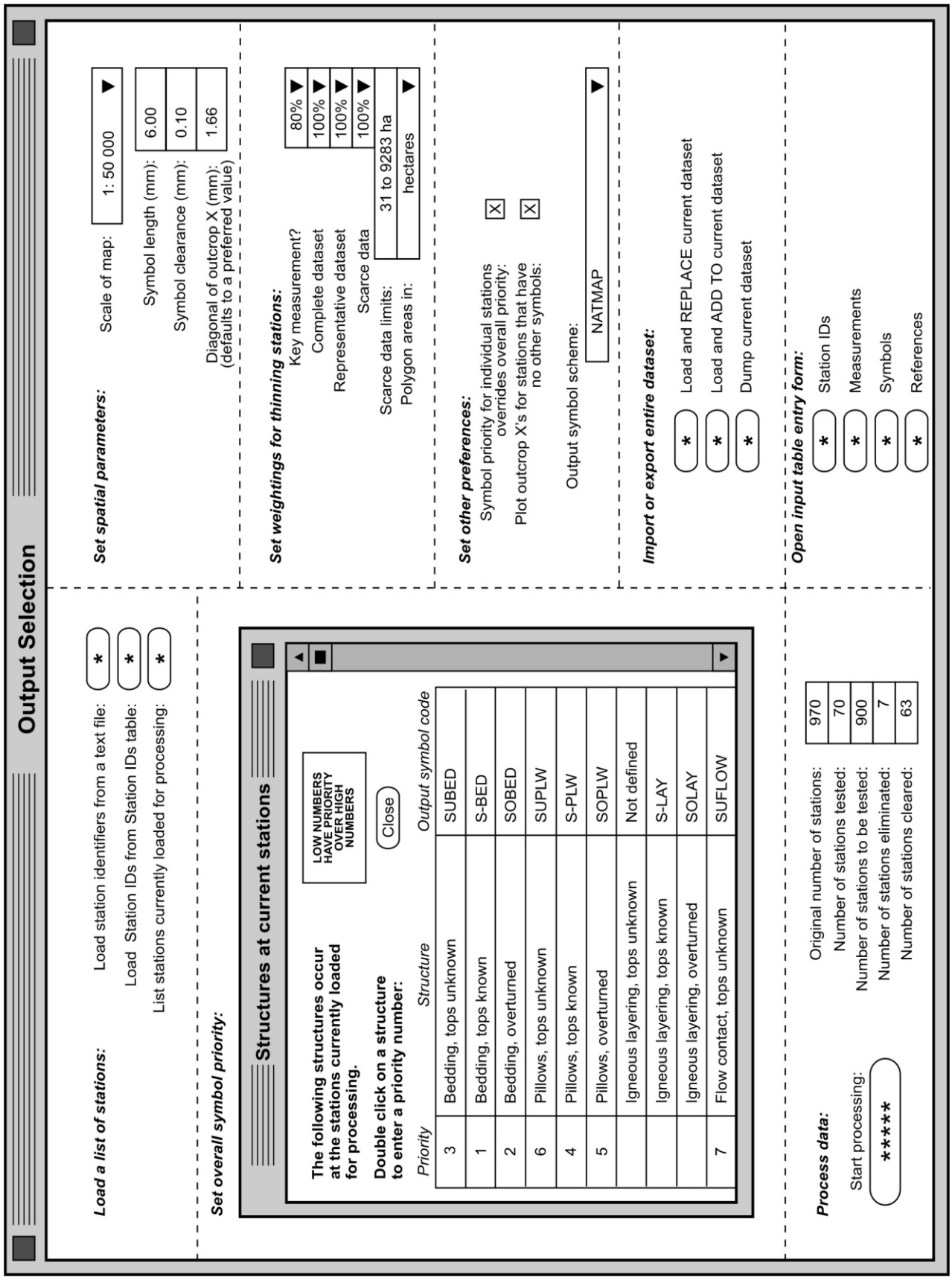

Figure 4. *Caption on facing page.*

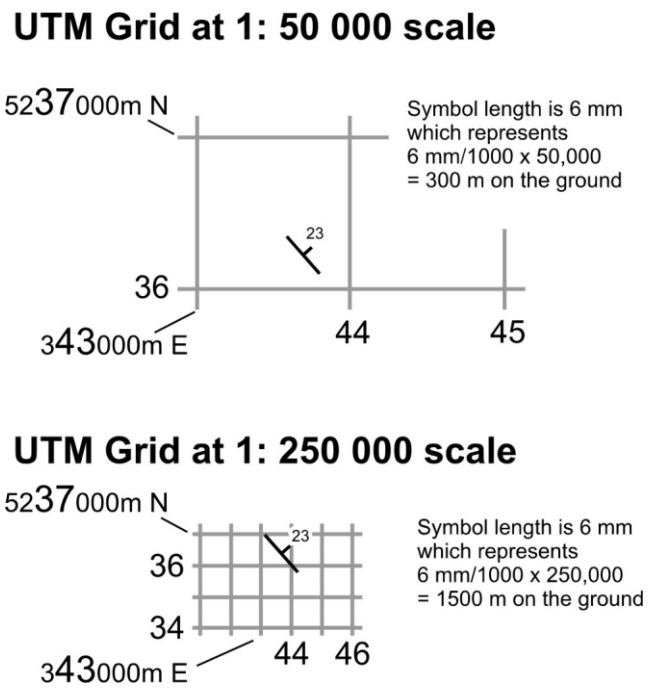

Figure 5. *Symbols on maps of different scales. In normal use, symbols stay the same size and the map base shrinks as scale is decreased.* GeoSymbol *takes an alternative view: the map base, expressed in metres on the UTM grid, is considered constant and the sizes of symbols, expressed in metres represented on the ground, increase.*

which a single measurement is considered to be representative. The limits are related to the space occupied by the symbol expressed in terms of the area this space represents on the ground. *GeoSymbol* calculates default limits, which should suffice under most circumstances, but to do this it requires that the user enter, on the 'Output Selection' form, the units being used in the database to express the areas of polygons.

Other Preferences

Two boxes are checked by default. The first one affects symbol priority. If it is checked, the symbol priorities entered for individual stations in the 'Measurements' table take precedence over the priorities entered for the overall map. If it is unchecked, the overall priorities take precedence. Normally only a few stations have individual priority entries. When the second box is checked, stations that have no structural measurements are marked by outcrop symbols, provided there is sufficient space on the map. If it is unchecked, these stations are omitted and the only stations plotted are those with structural symbols.

The 'Output symbol scheme' field allows the user to generate codes in the output file that match one of a variety of different symbol sets. The structural data in *GeoSymbol* can be entered using any symbol coding scheme. The codes must be defined in the 'Symbols' table in terms of the structures they represent and the schemes they belong to. Thus two records in the 'Measurements' table may contain the codes, 'S-FOL1' and 'ifol1', from the NATMAP and Ontario Geological Survey (OGS) symbol schemes, respectively (Figure 6). Both codes are defined as 'Foliation', '1st generation' in the 'Symbols' table so *GeoSymbol* treats them as representing the same structure. If both the NATMAP and

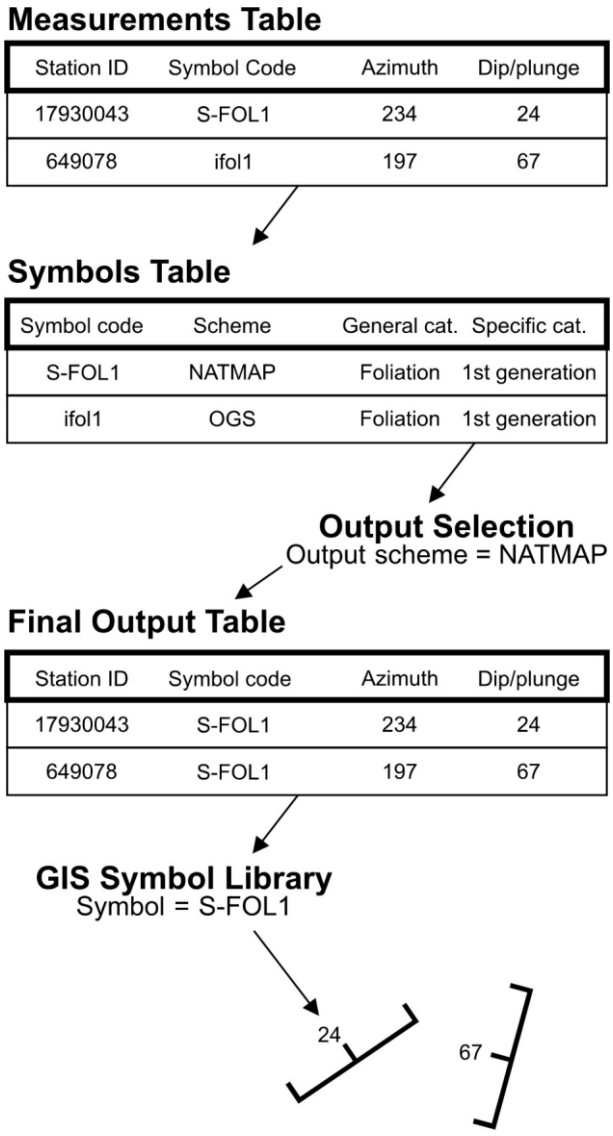

Figure 6. *Translation of symbol codes. Codes from any coding scheme can be entered into the 'Measurements' table. Codes are translated into general and specific category names, which are used for all internal processing. Any coding scheme that has been defined in the 'Symbols' table can be used for output, allowing easy access to a variety of symbol sets.*

OGS symbol schemes have been defined, both will be available on the ʻOutput symbol schemeʼ menu. If NATMAP is selected, all first-generation foliation measurements, no matter how they were entered in the ʻMeasurementsʼ table, produce a symbol code ʻS-FOL1ʼ on the output file. When the file is loaded into the graphical component of the GIS, the code causes the NATMAP symbol to be plotted on the map. If OGS is selected as the output scheme instead, the code produced is ʻifol1ʼ, which accesses the equivalent OGS symbol.

Final Changes and Plotting Formats

After the various parameters and preferences have been set, the user clicks the ʻStart processingʼ button and *GeoSymbol* proceeds to gather from its input tables all the information it needs to plot a symbol layer for the required map. It makes a set of temporary processing tables and then presents a dialog box to the user, which allows two things to be done.

First, the user can view individual stations and make temporary changes to the ʻSymbol priorityʼ, ʻKey measurement?ʼ and ʻSuppress measurement?ʼ fields in the ʻMeasurementsʼ table. Changes made at this stage only affect the current output file and are not saved in the main ʻMeasurementsʼ table. They provide a further opportunity to customize the map if needed.

Second, the user can now start the final processing by selecting the format for plotting symbols from the three choices presented in the dialog box (Figure 7). The map may be plotted with one symbol at each station, plotted directly on the station location. The second choice is to plot one symbol at each station, but to mark the actual location with an outcrop symbol and to plot the structural symbol off to one side. Alternatively, the user can choose to mark the actual location with an outcrop symbol and plot up to four structural symbols around it. Stations that have no structural symbols are simply marked by an outcrop symbol unless that option has been turned off.

Output File

The output file generated by *GeoSymbol* is intended to be loaded directly into the graphical component of the GIS and plotted "as is". For each symbol, including outcrop symbols, the file gives the ʻStation IDʼ, the UTM location, the ʻSymbol codeʼ, and the ʻAzimuthʼ and ʻDip/plungeʼ, if applicable. The UTM coordinates are used in this context to place each symbol in the correct position on the map and they are not necessarily the coordinates of the station itself. For example, if four structural symbols are being plotted at a particular station, there will be five records on the output file (Figure 8). One will have the ʻSymbol codeʼ for an outcrop

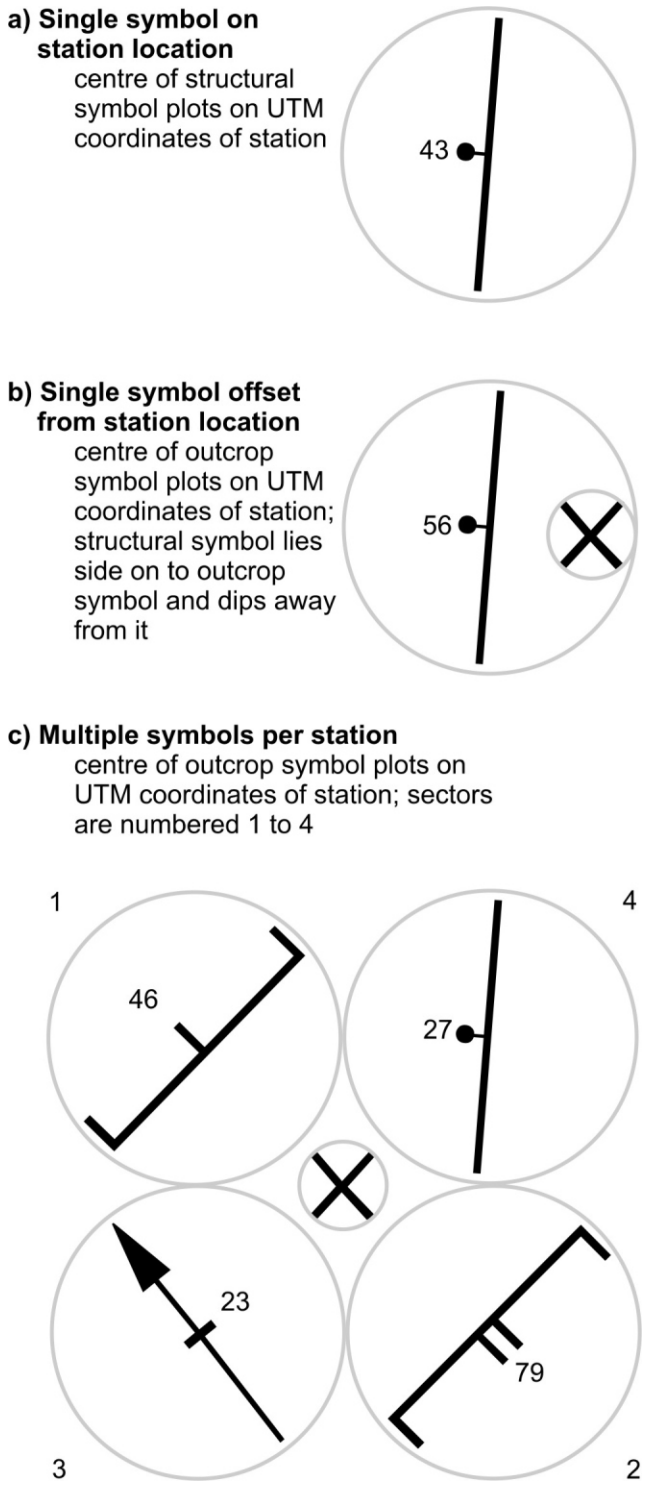

Figure 7. *The three plotting formats for final output.*

symbol and will have the UTM coordinates of the station; it will cause an outcrop symbol to be plotted right on the station location. The other four records will all have UTM coordinates that are different from those of the station location and different from each other. Their coordinates will

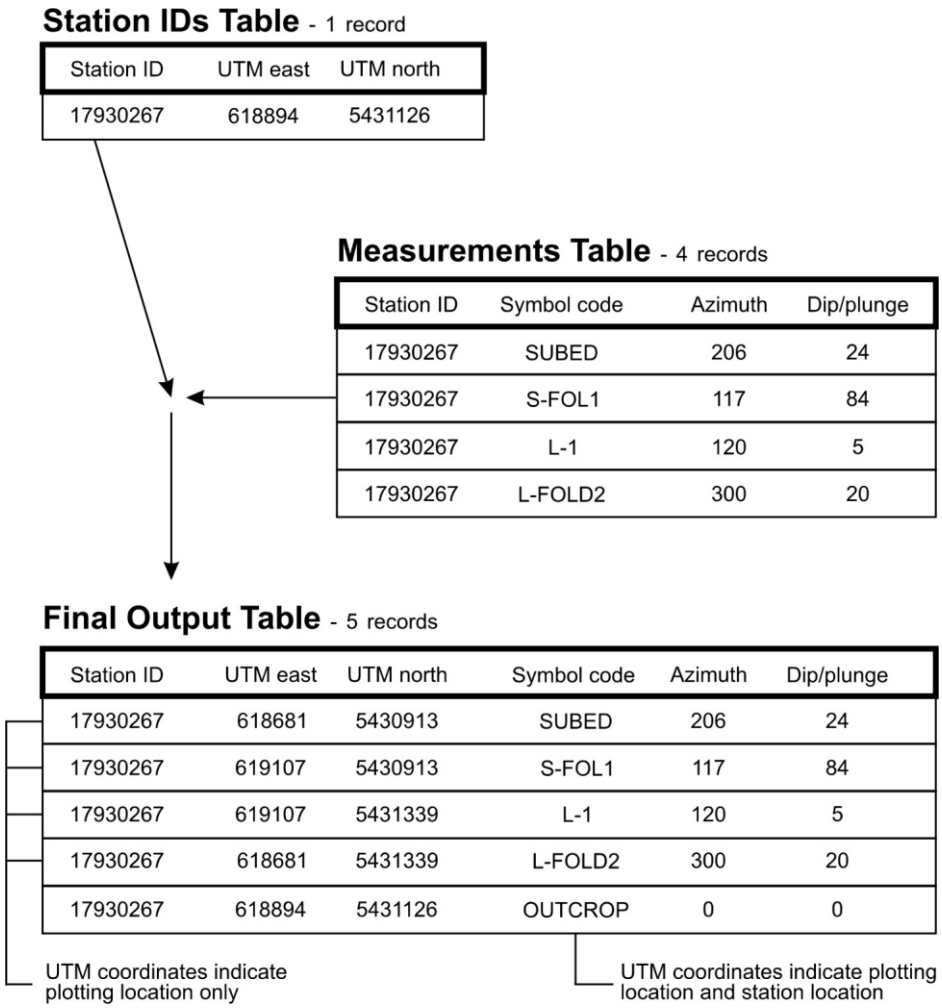

Figure 8. *Example of the relationship between the input tables and the 'Final output' table, which generates the output file. The 'Station IDs' table has one record for each station; the 'Measurements' table can have many records per station, or none at all; the 'Final output' table may have between 1 and 5 records per station depending on the plotting format and the number of measurements at the station.*

cause the structural symbols to plot around the outcrop symbol, with offset distances determined by the scale of the map and the dimensions of the symbols as entered on the 'Output Selection' form.

HARDWARE AND SOFTWARE

The present version of *GeoSymbol* is a prototype running on a Power Macintosh™ computer using the Helix Express™ database package. It interfaces with other computer programs and platforms by importing and exporting data as text files. This mode of operation, as well as the currently slow processing speed, dictates batch processing rather than interactivity. However, the intention is that, like *GeoLegend* which it complements, it will be optimized for interactive use as part of an integrated GIS package. As such, it would provide one layer in a comprehensive, multi-layered, geoscientific database accessible through user-friendly GIS viewing software.

STRUCTURE OF THE DATABASE

Four main input tables and four subsidiary tables are used to enter and store data in *GeoSymbol*. Fifteen tables are used to process and output data but only four of these are visible to the user; the other eleven operate transparently in the background. Figure 9 shows the relationships between the input tables and the principal output tables.

INPUT TABLES

Station IDs

Figure 1 shows the entry form presented to the user for the input of station location data and related information

about map polygons. Data can be entered directly into all fields except those containing computed variables, but the most efficient method is to load a previously compiled text file. Such a text file combines location information, which in many cases will have been entered into Fieldlog from a digitizing tablet, and polygon information generated by the graphical component of the GIS.

Variables

'Station ID' is the identifying number for a field station and uniquely defines a record in the 'Station IDs' table. If structural measurements have been recorded at a station, this variable links to one or more records in the 'Measurements' table.

'UTM easting', 'UTM northing' and 'UTM zone' give the station location.

'Map ID' is the number assigned to the associated published geological map in Geological Survey Geofiles. A station may not actually appear on the published map for lack of room or other reasons but it is part of the dataset that includes the stations that are shown on the map.

'GIS ID' is a computed variable derived from the 'References' table. It is an abbreviated identifier for the map and is used in the graphical component of the GIS because 'Map ID' commonly contains too many characters.

'Map reference' is a computed variable derived from the 'References' table. It gives the short-form reference for the associated published geological map indicated by 'Map ID'.

'Polygon ID' is the number of the polygon within the map designated by 'Map ID' in which the station is located. 'Polygon ID' and 'Map ID' correspond to equivalent variables in the 'Polygon IDs' table of *GeoLegend* (Colman-Sadd, 1998).

'Polygon area' is the area of the polygon identified by the combination of 'Map ID' and 'Polygon ID'. The area may be expressed in square kilometres, hectares, ares or square metres, provided that the chosen unit is used consistently for all records and is entered on the 'Output Selection' form.

Command Buttons

'Open input table'. Clicking on one of these buttons opens the entry form for that table.

'Go to "Output Selection"' opens the 'Output Selection' form, which contains all the command buttons and menus required to generate the symbol output file for a map.

'Check for duplicate records' checks that there are no records with duplicate values of 'Station ID'.

'Load text file' opens a dialog box so that the user can select a text file to load into the 'Station IDs' table; when the selection has been made, the button loads the file. The text file must contain a field for each of the seven entry fields on the form, in the order that they appear on the form. The 'Station ID' and three UTM fields must be filled. If the other fields are left empty, the 'Representative data' and 'Scarce data' criteria for thinning stations will only work if polygon data are entered individually for each map through the 'Output Selection' form (*see* "Station Selection").

'Select records' opens a dialog box so that the user can select records with particular values of variables in the 'Station IDs' table.

'Print station list' prints a list of records that meet the criteria defined in the 'Select records' dialog box.

'List records' lists all records that meet the criteria defined in the 'Select records' dialog box.

The 'Open…', 'Go to…', 'Select…', 'Print…' and 'List…' buttons are present on most of the input and output forms and allow easy access and manipulation of the information in *GeoSymbol*.

Measurements

Figure 2 shows the entry form presented to the user for the input of structural measurement data. As in the 'Station IDs' table, the most efficient entry method is to load a previously compiled text file. In most cases, the text file can be directly downloaded from the equivalent table in Fieldlog, and then provided with three extra fields for entering the special information needed by *GeoSymbol*.

Variables

'Map ID', 'GIS ID' and 'Map reference' are all computed variables and are the same for a given 'Station ID' as in the 'Station IDs' table.

'Station ID' is the identifying number for the field station at which a structure was measured.

'Symbol code' is the code for the type of structure measured. Codes can be used from any coding scheme provided their meaning in geological terms is defined in the 'Symbols' table.

'Symbol scheme' is a computed variable giving the scheme to which 'Symbol code' belongs. If the code has not

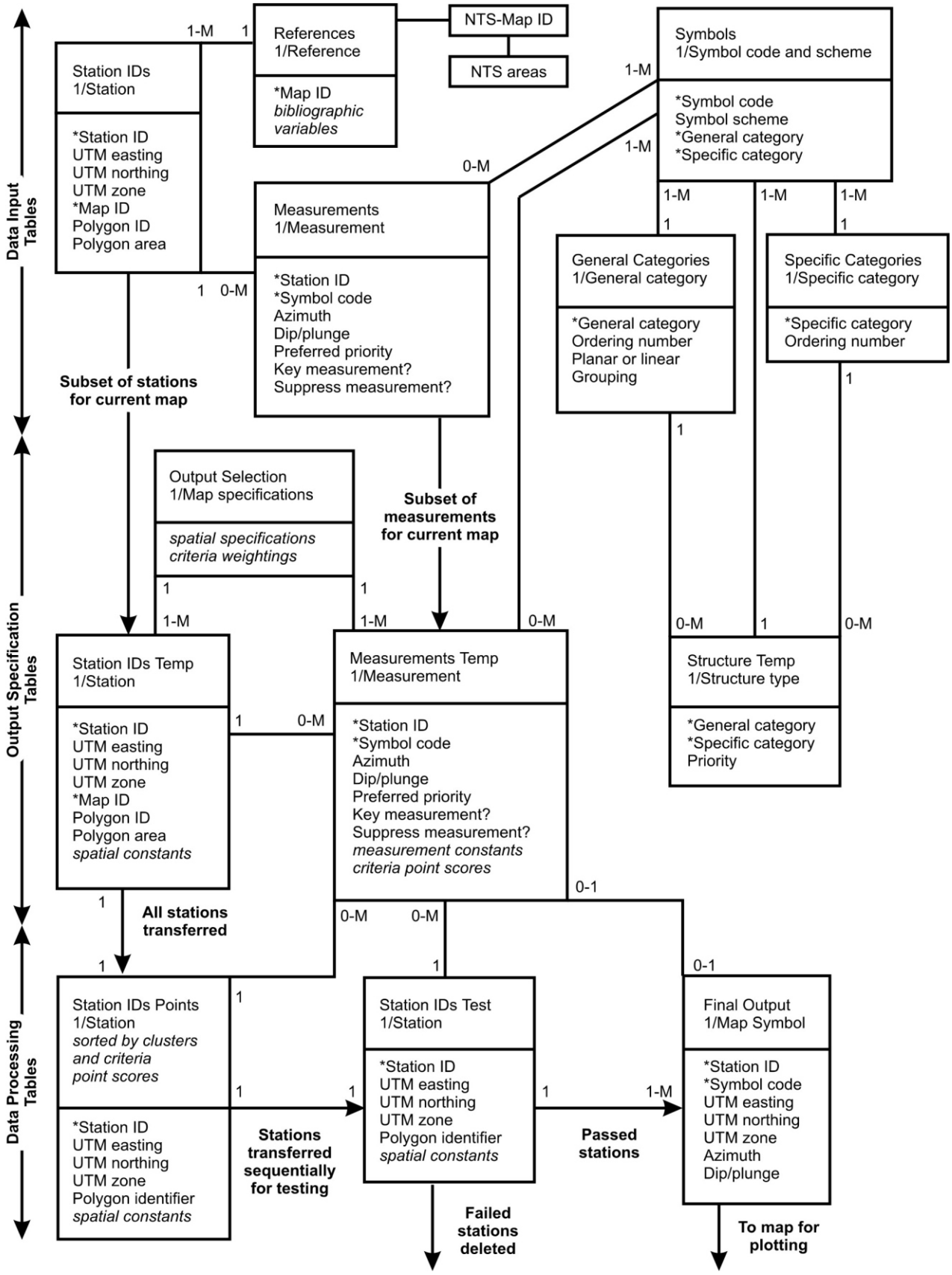

Figure 9. *Caption on facing page.*

Figure 9. *Organization of the input and principal output tables in* GeoSymbol. *Each box represents a table. The upper part of a box contains the table name and a description of a unique record, and the lower part lists the main fields (groups of fields are shown in italics). Asterisks indicate key fields. The lines joining boxes are links, and labels indicate the nature of the links (1, one record; 1-M, one to many; 0-M, zero to many; 0-1, zero to one).*

been defined, the field displays the message "Define code in Symbols table".

'General category' and 'Specific category' are computed variables derived from the 'Symbols' table. A 'Symbol code' always belongs to a 'General category' of geological features such as "Bedding" or "Foliation". It may also belong to a 'Specific category', which is always a subdivision of a 'General category', e.g., "Tops known" or "First generation".

'Azimuth' is the strike or trend of the structure, determined using the "right-hand rule".

'Dip/plunge' is the dip or plunge of the structure.

'Preferred priority' is the order in which measurements will be considered for plotting in the space available at the station. Lower numbers take preference over higher numbers. If the 'Symbol priority…' override on the 'Output Selection' form is unchecked, the values in 'Preferred priority' are ignored and the overall priority set on the 'Output Selection' form is used instead. Structures that have not been given an overall priority are omitted from a map, whether or not they have entries in 'Preferred priority'.

For most records, 'Preferred priority' is left blank and *GeoSymbol* defaults to the overall priority. 'Preferred priority' can be changed temporarily, for individual processing runs during the final editing stage, without affecting the values in the 'Measurements' table.

An example of the use of 'Preferred priority' is provided by a station where there are open folds of bedding and an axial planar foliation. The geologist has made four measurements of bedding, one of a fold axis, and one of the foliation. Suppose that the plotting format will place up to four symbols around each station and bedding has been given top overall priority. At this station, four bedding symbols will be plotted to the exclusion of the fold axis and foliation symbols. The 'Preferred priority' variable can be used to give two of the bedding measurements lower priority than the fold axis and foliation, allowing symbols for these structures to be plotted. If no overall priority is entered for fold axis and foliation, these structures are omitted and all four spaces will be occupied by bedding symbols.

'Key measurement?' can be checked to indicate a measurement that is of critical importance for some reason and is used by the criteria for thinning stations. Provided that the type of structure has been given an overall priority and will therefore be included on the map, the 'Key measurement?' designation gives the affected station preference in the thinning process relative to other stations that do not have key measurements. The 'Key measurement?' designation can be temporarily edited before final processing, if desired. Its effect can also be reduced or turned off by adjusting the weighting on the 'Output Selection' form. It should be noted that the preference given by this variable is relative so that checking a measurement at every station has the same effect as checking none at all or turning the criterion off by setting the weighting to zero.

'Suppress measurement?' can be checked to prevent a measurement from being included on a map. Some measurements may be of doubtful quality, perhaps because of uncertainty as to whether an outcrop is actually *in situ.* This variable allows such measurements to be retained in the database, as a matter of record, but prevents them from being plotted on a map until their validity has been confirmed. The 'Suppress measurement?' designation can be temporarily edited before final processing, an option that allows the user to remove individual measurements from a map.

Command Buttons

'Check for duplicate records' finds records in which all four of the variables 'Station ID', 'Symbol code', 'Azimuth' and 'Dip/plunge' contain identical values to those in another record.

'Load text file' opens a dialog box so that the user can select a text file to load into the 'Measurements' table; when the selection has been made, the button loads the file. The text file must contain a field for each of the seven entry fields on the form, in the order that they appear on the form. The 'Station ID' and 'Symbol code' fields must be filled. The other fields may be left empty but most structures require at least an azimuth to plot satisfactorily on a map.

Other command buttons perform similar functions to those on the 'Station IDs' entry form.

Symbols

Figure 3 shows the entry form presented to the user for defining symbol codes. Structural symbols usually incorpo-

rate a number of separate concepts. In order for a computer to make geologically meaningful decisions about symbols, these concepts have to be separated from each other. This is already done to some extent in a typical Fieldlog file, where 'Azimuth' and 'Dip/plunge' are placed into separate fields from 'Symbol code'. However, the process is not complete and 'Symbol code' still amalgamates features that need to be distinguished. For example, in the NATMAP scheme, "B-Z1" indicates a first-deformation Z-fold, providing both time of formation and a geometrical description. In terms of geometry, "B-Z1" is more comparable to "B-Z2", a second-deformation Z-fold, than it is to "B-S1", a first-deformation S-fold. However, in terms of timing, and hence the geological events that produced the structures, "B-Z1" is more closely related to "B-S1". If a user wants to select symbols that are representative of the orientation of first deformation fold axes, it is geologically meaningful to compare "B-Z1" and "B-S1", and geologically meaningless to compare "B-Z1" and "B-Z2". The two variables, 'General category' and 'Specific category', separate the disparate concepts incorporated into 'Symbol code' so that stations can be thinned using geological criteria. *GeoSymbol* uses these two variables, rather than 'Symbol code', to identify structures in all internal operations.

Variables

'General category' gives the general geological feature indicated by 'Symbol code'. The general category would normally include such terms as "Bedding", "Igneous layering" or "Foliation". 'General category' must be defined for all records. For some purposes, it is necessary to group different types of 'General category' together into even broader geological concepts; this is discussed in the "Command Buttons" section.

'Specific category' is an optional subdivision of 'General category'. For example, "Bedding" may be divided into "Tops known", "Overturned" and "Tops unknown" or may have even more detailed specific categories like "Tops known from cross bedding" or "… from grain gradation". Although the latter could be considered a subdivision of the former, an extra level in the hierarchy would not serve any useful purpose in *GeoSymbol* so all of these variations are placed at the same level. Deformational episodes for secondary structures, such as "First generation" and "Second generation", are also placed in 'Specific category'. Note that differences in dip or plunge, indicated by "horizontal", "vertical" and "inclined", are not considered as categories even though many symbol schemes use separate symbols for them; *GeoSymbol* assumes that only the inclined symbol will be used and the attitude will be shown by numbers from 0 to 90. Both 'General category' and 'Specific category' must be filled from pop-up menus; this ensures typographic consistency and the entry of important associated data (*see* "Command Buttons").

'Symbol scheme' is the name of the scheme to which a 'Symbol code' belongs, e.g., "SUBED" is the code for "Bedding, tops unknown" in the "NATMAP" scheme. If the same code is used to indicate the same geological structure in two different schemes, and both of these schemes are to be used for output, the code must be entered twice, once for each scheme. If the same code is used to indicate two different structures in two different schemes, one of the codes must be changed, perhaps by adding a suffix, so that the two can be distinguished.

'Symbol code' is the code used under a particular symbol scheme to indicate a type of structure. A record in the 'Symbols' table is uniquely identified by a combination of 'Symbol scheme' and 'Symbol code'.

Command Buttons

'Define a new category: General' allows new items to be made available on the 'General category' menu of the 'Symbols' entry form. Clicking the command button opens a form that lists all the current menu items and shows an entry field for typing in new ones (Figure 10). There are three other fields on the form.

'Ordering number' controls the order in which items appear on the menu, with "1" placed at the top and increasing numbers downward.

'Planar or linear' must be entered to describe the geometric nature of the structure. The 'Representative data' criterion for thinning stations compares measurements by plotting the poles to planar structures or the actual linear structures on a "virtual" stereogram. In order to do this, it must be told whether a structure is planar or linear and, therefore, whether it should plot the pole or the structure itself.

'Grouping' provides an optional higher level in the categorizing of 'Symbol code'. For example, several kinds of 'General category' ("Bedding", "Flow contact", "Pillows") indicate the paleohorizontal when the rocks were formed. If all of these are designated "Paleohorizontal" in the 'Groupings' field, they are all compared together when *GeoSymbol* determines the most representative measurements in a polygon. Likewise, there are many different kinds of folds (S-folds, Z-folds etc.), but in order to choose representative fold-axis symbols, we may want to compare all of these fold types together. Placing the entry "Fold axis" in the 'Grouping' field achieves this. There is an important difference between primary and secondary structures. Primary structures are known to have all formed at the time of for-

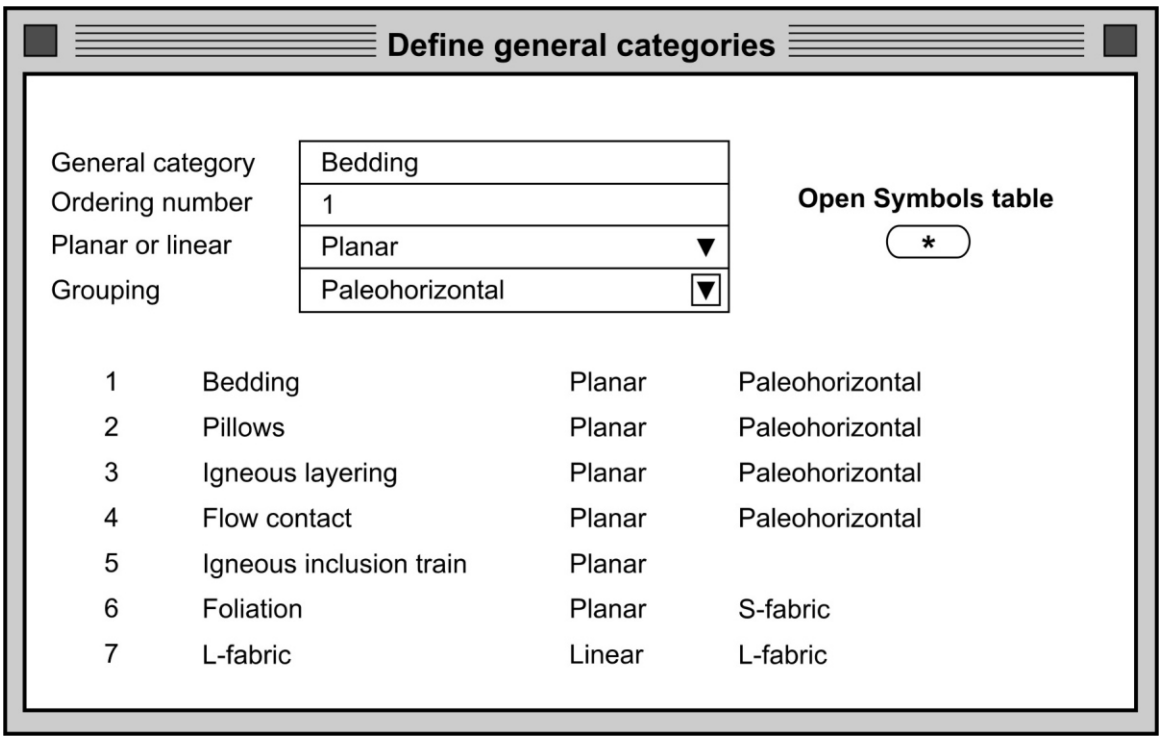

Figure 10. *Entry form for the 'General categories' table.*

mation of the rock and the entry in 'Specific category' is used for some feature that is not normally relevant to the grouping. Secondary structures, however, may have formed at quite different times and be unrelated in terms of genesis; for these structures, 'Specific category' is normally used to denote the deformational episode and is critical to how structures are grouped. Therefore, if 'Grouping' starts with the key words "Paleohorizontal", "Paleocurrent" or "Other primary structure", *GeoSymbol* ignores entries in 'Specific category' for the purpose of comparing measurements. In all other cases, it assumes that the structure is secondary and that any entry in 'Specific category' refers to the deformational episode. It only compares measurements of these structures if they have identical entries in the 'Specific category' field.

'Define a new category: Specific' allows new items to be made available on the 'Specific category' menu of the 'Symbols' entry form. Clicking the command button opens a form that lists all the current menu items and shows an entry field for typing in new ones. The form also has an 'Ordering number' field to control the sorting of items on the menu.

References

The 'References' table contains all the information required to make a short-form or full bibliographic reference for a map. References should be entered for every map that occurs in the 'Map ID' field in the 'Station IDs' table. This field links to an equivalent field in the 'References' table so that the source of every station and every measurement is recorded. Two subsidiary tables relate individual references to their respective NTS areas. The 'References' table is the same as that used in *GeoLegend* and a full description is provided in the manual for that software (Colman-Sadd, 1998).

OUTPUT PROCESS

Station Selection

Stations that are to be considered for plotting on a map are selected either by loading a text file into *GeoSymbol* or by running a query on the contents of the 'Station IDs' table. *GeoSymbol* creates a temporary table, 'Station IDs Temp' (Figure 9), to hold these stations during processing and to calculate various spatial constants. The 'Station IDs Temp' table is only visible to the user as a listing within the 'Output Selection' form.

Using a Text File

A text file would normally be used if a map is required for a specific area. The area is outlined in the graphical component of the GIS and a list of all stations contained within it is downloaded. The list is formatted as a text file with three fields for each record. The first field contains the 'Station ID' and must be filled. The second field is reserved

for the unique identifier of the map polygon in which a station occurs, and the third field for the area of that polygon in whatever units have been selected for use in the database as a whole (*see* "Station IDs"). The second and third fields can be left empty, in which case *GeoSymbol* fills them automatically with default values from the 'Station IDs' table. The default value for the unique polygon identifier is a combination of 'GIS ID' and 'Polygon ID', in the form NF045_0067, and corresponds to the polygon identifiers used in *GeoLegend*. The default value for the polygon area is derived from the 'Polygon area' field.

If values are entered into the fields for polygon identifier and polygon area in the text file, these values take precedence over those stored in the 'Station IDs' table. The option to include polygon information with the text file allows thinning criteria to be calculated using various polygon structures for a single map. It is intended that the default polygon information entered in the 'Station IDs' table will be derived from maps showing units divided at their maximum level of detail. Different polygon structures may be produced if *GeoLegend* is used to generalize or customize maps. If a map is to be plotted showing units generalized to the group level, the graphical component of the GIS merges all adjacent polygons belonging to the same stratigraphic group to produce fewer, larger polygons. If data for these polygons are then loaded into *GeoSymbol*, thinning criteria are calculated by comparing all measurements of a particular kind within each of the larger, merged polygons, rather than limiting comparisons within the smaller polygons of the detailed map.

When the text file has been assembled in the correct format, it is loaded into *GeoSymbol* by clicking 'Load station identifiers from a text file' on the 'Output Selection' form (Figure 4) and selecting the file from a dialog box. The previous contents of the 'Station IDs Temp' table are deleted and the new data are loaded in their place.

Using the 'Station IDs' Table

Stations can also be selected directly from the 'Station IDs' table by clicking 'Load Station IDs from Station IDs table' (Figure 4). A selection form allows different values to be set for the variables in the table and all stations that meet the specified criteria are loaded into the 'Station IDs Temp' table after the previous contents have been deleted. This method can be used to select stations for a specific area, using UTM coordinates, and then to select a subset of these stations by specifying limits for the 'Station ID' variable.

'Station IDs Temp' Table

The 'Station IDs Temp' table is used to assemble the location information required for processing the data and to calculate various spatial constants. Stations in the table can be viewed by clicking the 'List stations currently loaded…' button (Figure 4). Apart from 'Station ID', the list shows the polygon information that may or may not have been loaded with the text file, UTM information and 'Map ID', derived from the 'Station IDs' table, and reference information, derived from the 'References' table. Any of the stations can be deleted, if they are not needed for the map, without affecting the contents of the 'Station IDs' table.

Measurement Selection

Two temporary tables are used to assemble the structural measurement data required for a map (Figure 9). The 'Structure Temp' table is used to choose and prioritize the different types of structures that are to be plotted. The actual data are collected in the 'Measurements Temp' table, where various constants related to structural measurements are calculated.

Setting Overall Symbol Priority

The 'Structures at current stations' window on the 'Output Selection' form lists all the records in the 'Structure Temp' table (Figure 4). The list shows all the types of structures represented at the stations in the 'Station IDs Temp' table. It is derived by referring the values of 'Station ID' to the 'Measurements' table, identifying all the 'Symbol codes' at these stations and then referring these to the 'Symbols' table to derive the general and specific category names for each code. The category names are concatenated to provide the structure names used by *GeoSymbol*, and these names are listed in the centre column of the list, e.g., "Igneous layering, tops known".

If data have been entered into the 'Measurements' table using different symbol schemes, it is possible that several different values of 'Symbol code' in that table may be used for the same type of structure. In this case, each of these codes will be defined in terms of the same general and specific categories. These, in turn, will give rise to a single structure name on the list of records in the 'Structure Temp' table.

The right column in the list gives the 'Symbol code' for the structure using whatever symbol scheme is currently shown in the 'Output symbol scheme' field of the 'Output Selection' form. This code is the one that will be used to represent the structure on the output file that is sent to the graphical component of the GIS. If no code has been defined for a given structure under the selected output scheme, the right column shows "Not defined". In this case, either the code must be defined in terms of general and specific categories in the 'Symbols' table or a different symbol scheme must be selected for the output.

The relative priority for structures is set in the left column. A number is entered for all structures that are to be plotted on the map. If a particular kind of structure is to be completely omitted from the map, this is achieved by leaving the priority column blank for that structure. The order of priority determines the preference given to different symbols at individual stations. For example, it may have been decided to plot bedding and first-foliation symbols for a map and only to plot one symbol at each station. Only bedding and first foliation are given priority numbers, thus suppressing all other symbols. The bedding symbols may be numbered 1, 2 and 3, for "tops known", "overturned" and "tops unknown", respectively, and first-foliation may be numbered 4. If a bedding measurement occurs at a station, this is plotted and first-foliation measurements are omitted because only one symbol per station is specified. If there are no bedding measurements, then a first foliation symbol can be plotted in the available space. Similarly, if there are "tops known" and "tops unknown" bedding measurements at the same station, the "tops known" measurement is plotted. If the format that allows up to four symbols per station has been chosen, *GeoSymbol* uses the priority list to decide which symbols in excess of four will be omitted from a station.

There may be some stations that require a special priority order for their symbols. This order can be set using the 'Preferred priority' field, either in the 'Measurements' table, where it will apply every time the data are used to make a map, or in the 'Measurements Temp' table (*see below*), where it only applies to the current version of the map. An example of the use of this field is given in the variable descriptions for the 'Measurements' table, above. If no special order is required, the 'Preferred priority' field is left blank and *GeoSymbol* uses the overall priority entered on the 'Structure Temp' table. Preferred priorities can be suppressed, so that all stations follow the overall priority, by unchecking the 'Symbol priority for individual stations…' box on the 'Output Selection' form (Figure 4). The default is for the box to be checked.

The second box in the 'Set other preferences' section allows outcrop symbols to be suppressed at stations that have no structural measurements. These stations include, first, those that have no measurements recorded in the 'Measurements' table and, second, those that do have measurements but not of a type selected for plotting by the entry of an overall priority number.

'Measurements Temp' Table

When the 'Start processing' command button is clicked on the 'Output Selection' form, the information entered for overall priority is used to create a subset of measurements for the map. These measurements are stored in the 'Measurements Temp' table where they can be edited for the particular map being produced without affecting the data in the 'Measurements' table. It is possible to delete complete records or to change values in the 'Preferred priority', 'Key measurement?' or 'Suppress measurement?' fields. The editing option is available as a command button on the same form that allows the plotting format to be selected. The form is presented to the user as soon as the data in the 'Measurements Temp' table have been assembled.

Station Geometry

The second part of the form for choosing plotting formats consists of three command buttons. Any one of these buttons may be clicked to set in progress the main processing of the data and the output of a file ready for loading into the graphical component of the GIS. The button that is clicked determines the plotting format of symbols on the map (Figure 7). Before the formats are described, it is necessary to define the two symbol radii that are fundamental to all spatial calculations involved in the plot.

Symbol Radii

Each structural symbol is allotted a circular space on the map (Figure 11). The radius of the circle, R, is expressed in metres represented on the ground and therefore uses the same units as the UTM grid. Similarly, each outcrop symbol ('x') is allotted a circular space with radius r, also expressed in metres represented on the ground. The two radii are determined as follows:

$$R = (s / 1000)(l + 2c) / 2$$

$$r = (s / 1000)(d + 2c) / 2$$

where s is the second term of the scale ratio (e.g., 50,000);
l is the symbol length in millimetres;
c is the symbol clearance in millimetres;
d is the diagonal of the outcrop symbol in millimetres.

The variables, s, l, c and d, are derived from entries in the spatial parameters section of the 'Output Selection' form. The factoring of the map scale, s, into R and r automatically adjusts all calculations of offsets and overlaps for the scale of the map.

The diagonal of the outcrop symbol has a default value, $d$def, and a maximum value, $d$max, on the 'Output Selection' form, both of which are functions of l and c:

$$d\text{def} = (l + 2c) / 2 - (l + 2c) / 5 - 2c$$

$$d\text{max} = \sqrt{2\left((l+2c)/2\right)^2} - (l + 2c)/2$$

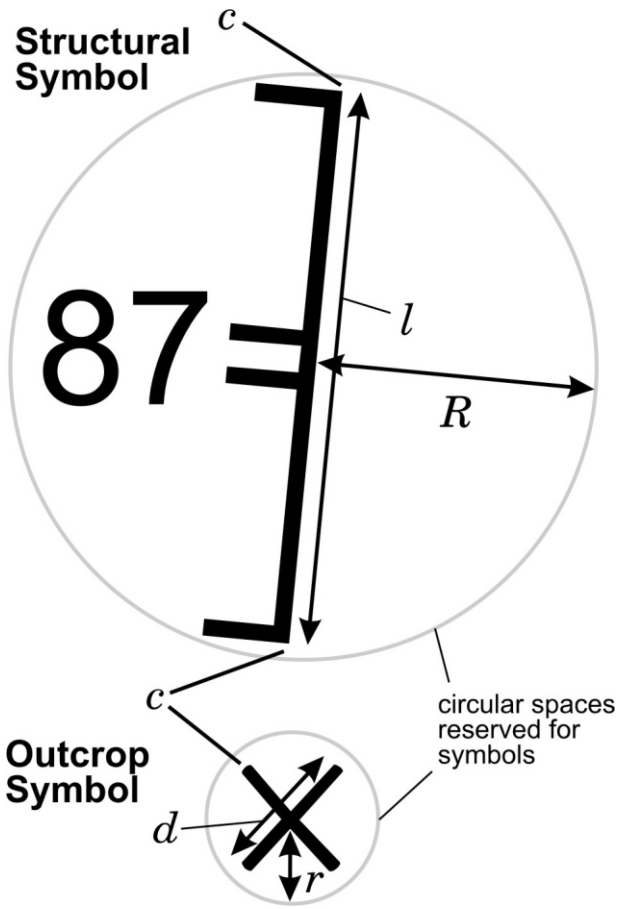

Figure 11. *Variables used to define symbol dimensions. R, radius of space reserved for a structural symbol; l, length of a structural symbol; r, radius of space reserved for an outcrop symbol; d, diameter of an outcrop symbol; c, clearance at the end of each symbol. See text for units used for each variable.*

The default value, which is automatically calculated if the entry field is cleared, is the largest size that will fit on the map without overlap using any of the possible printing formats. The maximum value is the largest size that will fit without overlap using the option to plot multiple symbols at a station; this size will, however, cause overlap if used with the single-symbol offset format.

Single Symbol On Station Location

This format plots one structural symbol at each station and centres the symbol at the station location, as defined by the UTM coordinates (Figure 7a). The symbol plotted is the one that has the highest priority of those available for the station (*see* "Setting Overall Symbol Priority"). The space reserved for the symbol on the map is a circle with radius *R* centred on the station location.

If no structural symbols are available for a station and the option to plot outcrop symbols is checked, an outcrop symbol is plotted. It is centred on the UTM coordinates for the station and a circular area with radius *r* is reserved for it.

Single Symbol Offset from Station Location

Using this option, a single structural symbol and an outcrop symbol are plotted at each station. The outcrop symbol is plotted exactly at the UTM coordinates for the station. The structural symbol is plotted to one side in such a way that it lies side-on to the outcrop symbol and dips away from it (Figure 7b). A circular space with radius *R* is reserved for both symbols at the station and is centred on the offset centre of the structural symbol, not on the station location. The outcrop symbol is plotted within the circle reserved for the structural symbol. There will be no overlap if the geometrical specifications for the structural symbols have been observed and the outcrop symbol is no larger than the default value on the 'Output Selection' form.

The centre of the structural symbol is offset from the station location by a distance, D_1, that is expressed in metres represented on the ground. This distance is given by the relation:

$$D_1 = 2R / 5 + r$$

The position of the offset structural symbol is fixed by its own set of UTM coordinates, which are calculated from the coordinates for the station:

$$O_E = E + D_1(\cos t)$$

$$O_N = N + D_1(-\sin t)$$

where O_E is the easting for the offset structural symbol;
O_N is the northing for the offset structural symbol;
E is the easting of the station;
N is the northing of the station;
t is the 'Azimuth' of the structure from the 'Measurements' table.

As in the case of the single symbol plotted at the station location, if no structural symbols are available, an outcrop symbol is plotted at the UTM coordinates for the station within a circular space of radius *r*.

Multiple Symbols per Station

This option allows a maximum of four symbols to be plotted at a station. The station location itself is marked by an outcrop symbol within a circular area of radius *r*. The

structural symbols are plotted in four circular sectors, each of radius *R*, located northwest (Sector 1), southeast (2), southwest (3) and northeast (4) of the station location, relative to the UTM grid (Figure 7c). If there are more than four symbols at a station, the ones plotted are those with the highest priority. If there are fewer than four symbols, the vacant sectors can be overlapped by other stations. The assignment of symbols to particular sectors is optimized to allow the maximum number of stations to be plotted on the map without overlap. If there are no symbols at a station, an outcrop symbol is plotted within a circle of radius *r*, provided this option has not been turned off.

The UTM coordinates for the structural symbols are calculated by addition or subtraction:

Sector 1 (NW)	$M_{E(w)} = E - R$	$M_{N(n)} = N + R$
Sector 2 (SE)	$M_{E(e)} = E + R$	$M_{N(s)} = N - R$
Sector 3 (SW)	$M_{E(w)} = E - R$	$M_{N(s)} = N - R$
Sector 4 (NE)	$M_{E(e)} = E + R$	$M_{N(n)} = N + R$

where $M_{E(w)}$ is the easting for symbols in sectors west of the station;
$M_{E(e)}$ is the easting for symbols in sectors east of the station;
$M_{N(n)}$ is the northing for symbols in sectors north of the station;
$M_{N(s)}$ is the northing for symbols in sectors south of the station.

The offset distance, D_2, of each structural symbol from the station location is expressed in metres represented on the ground and is given by the relation:

$$D_2 = \sqrt{2R^2}$$

Conditions for Station Overlap

All structural symbols on the map are enclosed in invisible circles of radius *R*, and all outcrop symbols within circles of radius *r*. The centres of these circles are located by coordinates expressed in metres on the UTM grid; some are located at the station coordinates and some are offset as described above. Distances between circle centres are calculated in metres using coordinate geometry.

Using either of the single-symbol formats, the current station overlaps with another station if one of the following conditions is met:

$\delta_{s\to s}$ min $< 2R$
$\delta_{s\to o}$ min $< R + r$
$\delta_{o\to o}$ min $< 2r$
$\delta_{o\to s}$ min $< r + R$

where the distances between circle centres are expressed by
$\delta_{s\to s}$ if both the current station and another station have structural symbols;
$\delta_{s\to o}$ if the current station has a structural symbol but the other station does not;
$\delta_{o\to o}$ if neither the current station nor the other station has a structural symbol;
$\delta_{o\to s}$ if the current station does not have a structural symbol but the other station does.

If the multiple-symbol format is used, it is more complicated to determine whether two stations overlap. Each station having structural symbols contains five elements that must be checked for overlap individually. One of these is the circle of radius *r*, which contains the outcrop symbol, and the other four are circles of radius *R* for each of the sectors. *GeoSymbol* checks the minimum distance for each element to the elements of the other stations and registers overlaps if these distances are less than the allowed amounts, as given above. However, the overlap of a sector with an element of another station does not necessarily mean that the stations themselves overlap. This is because there may be fewer symbols at the station than there are sectors, and empty sectors do not generate overlaps (Figure 12). Symbols are assigned to the available sectors at a station so as to maximize the number of stations that can be plotted without overlap. A station is deemed to overlap another if there is an overlap of the circle around its outcrop symbol or if the number of structural symbols at the station exceeds the number of sectors that are free from overlap, after the distribution of the symbols has been optimized.

Calculation of Thinning Criteria

Only those stations that will fit on the map without overlap are plotted. Four criteria are used to determine the order of precedence for overlapping stations. These are calculated in the 'Measurements Temp' table where they result in a point score for each station. Except for the 'Scarce data' criterion, which has to be recalculated every time a station is dropped, the scores are calculated once, at the beginning of processing. 'Key measurement?' scores are expressed in thousands and are intended to create an entirely different class of stations that receive special treatment. Scores for the other criteria, with a minor exception in the case of 'Scarce data', are calculated out of 100. Scores are then multiplied by the weighting factors assigned on the 'Output Selection' form and are used to order the records in the 'Station IDs Points' table (Figure 9). This table contains the same stations as the edited version of the 'Station IDs Temp' table and is generated as a "working copy" for use during processing.

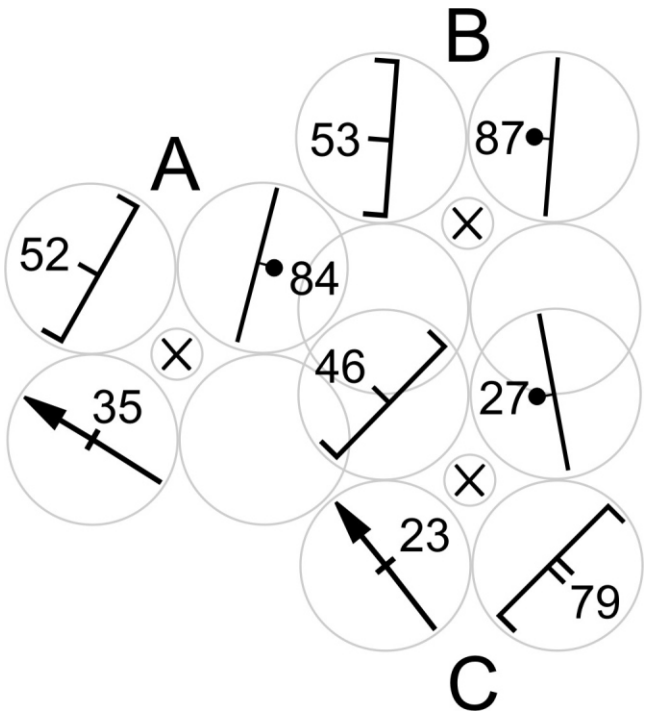

Figure 12. *Example of the optimization of symbol distribution in the multiple-symbol format. Symbols are assigned to sectors so that all three stations can be plotted without sectors containing symbols overlapping each other. Overlaps with empty sectors do not disqualify a station from plotting.*

Key Measurement?

Any measurement that is flagged as true in the 'Key measurement?' field is given a score, *K*, of 1000. For an unflagged measurement, *K* is 0.

Complete Dataset

Each measurement type must be given an overall symbol priority on the 'Output Selection' form in order for it to be able to plot on the map. The 'Complete dataset' criterion uses this priority number, p_{mnt}, to favour stations that have both more measurements and higher priority measurements. For a station with *n* measurements, the reciprocals of the overall priority numbers are summed, and the total, p_{stn}, is expressed as a percentage of the maximum value for any station on the map to give the 'Complete dataset' point score, *C*, for the station:

$$p_{stn} = \sum_{i=1}^{n} 1 / p_{mnt}$$

and

$$C = p_{stn}\ (100 / p_{stn}\ \max)$$

'Complete dataset' is the only criterion that places a collective value on all the data at a station. For this reason, *C* is identical for every measurement at a particular station.

Representative Data

The 'Representative data' criterion favours measurements that are most representative of particular groupings of structures within a polygon. *GeoSymbol* uses spherical trigonometry to compare measurements on a mathematical representation of a stereographic projection (Phillips, 1960). For this purpose, structures are expressed in linear form. Although intersections with both the upper and lower hemispheres must be considered, all necessary values can be derived from angles in the lower hemisphere (Figure 13).

For linear structures:
$\alpha = t$ and $\beta = 90 - e$

and for planar structures:
for $t \geq 90$ $\alpha = t - 90$
for $t < 90$ $\alpha = t + 270$
and $\beta = e$

where *t* and *e* are the 'Azimuth' and 'Dip/plunge', respectively, from the 'Measurements' table;
α is the azimuth of a linear structure or a pole to a planar structure, projected from the lower hemisphere;
β is the angle between the centre of the lower hemisphere and a linear structure or a pole to a planar structure in the lower hemisphere.

Every structure has two intersections with the sphere. The attitudes of structures are compared by calculating the minimum angular distances between them, if they are linear, or between their poles, if they are planar. This distance ranges from 0° and 90°. The upper hemisphere is required because the minimum angular distance between two structures may not be included solely within the lower hemisphere. For example, suppose that a number of measurements have been made on beds that strike approximately east–west and are close to vertical. Some of the beds may dip slightly to the north and some slightly to the south. The poles for those dipping north will cluster around the south pole of the lower hemisphere and the poles for those dipping south will cluster around the north pole. The angular distance between a measurement in the north cluster and a measurement in the south cluster, as calculated on the lower hemisphere, is close to 180°. However, each pole also has an intersection with the upper hemisphere. If the upper hemisphere is included, the minimum angular distance between any two poles is close to 0°, which is a true representation of the fact that all the measurements of bedding are very similar.

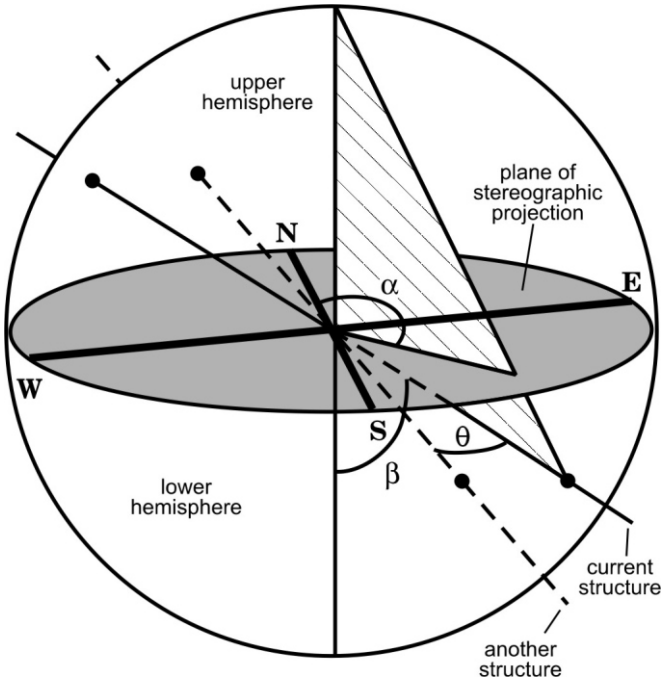

Figure 13. *Variables used in the calculation of the 'Representative data' criterion, depicted within the spherical construction used to generate a stereographic projection. The two structure lines represent linear structures or the poles to planar structures.*

For every measurement in a polygon, *GeoSymbol* calculates the minimum angular distance from the lower hemisphere intersection of the current record to every other comparable measurement in the same polygon, and then averages the distances. The most representative measurements have the lowest average distances. Angular distances are initially calculated in the lower hemisphere and these are used to derive the minimum angular distances from the absolute values of cosθ:

$$\cos\theta = \left(\cos\beta.\cos\beta^*\right) + \left(\sin\beta.\sin\beta^*.\cos\left|\alpha - \alpha^*\right|\right)$$

$$\phi = \tan^{-1}\frac{\sqrt{1-\cos\theta^2}}{\left|\cos\theta\right|}$$

where θ is the angular distance between the lower hemisphere intersection of the current measurement and the lower hemisphere intersection of another measurement (Figure 13);
values of α and β for the current measurement are signified by asterisks;
ϕ is the minimum angular distance calculated from the lower hemisphere intersection of the current measurement to either the lower or upper intersection of another measurement.

The 'Representative data' score, V, is the complement of the average minimum angular distance to comparable measurements in the polygon, expressed as a percentage of 90°:

$$V = \left(90 - \frac{\sum_{i=1}^{m}\phi}{m}\right)100/90$$

where m is the number of comparable measurements from other stations within the polygon.

The comparison of measurements is based on entries in the 'Grouping', 'General category' and 'Specific category' fields in the 'Symbols' and subsidiary tables. Measurements of primary structures are comparable if the structures have the same 'Grouping' or, if 'Grouping' is undefined, the same 'General category'; secondary structures must also have the same 'Specific category' because this is where the relative ages of the structures are indicated.

If a measurement is the only one of its kind in a polygon, V defaults to a score of 100. Not only is this logical, because the measurement must be completely representative of itself, but it also has the desirable effect of giving preferential treatment to the only source of a particular kind of structural information in the polygon and reinforcing the effect of the 'Scarce data' criterion.

Measurements that have incomplete information, usually because 'Dip/plunge' has not been entered, cannot be plotted on a stereogram and are given scores of zero.

Scarce Data

Scores of 'Scarce data', S, are calculated out of 100 with one exception. If a measurement is the last of its kind within a polygon, *i.e.*, m=1, it defaults to a score of 150, which together with a defaulted score of 100 for 'Representative data' is intended to provide the station with preferential treatment relative to its neighbours.

For other measurements, the score is calculated by dividing the polygon area by the number of comparable measurements in a polygon, and expressing this as a percentage of the maximum value of this quotient for the most common type of measurement on the map:

$$S = \left(\frac{A}{m}\right)\left(100\Big/\left(\frac{A}{\mu}\max\right)\right)$$

where A is the area of a polygon;
m is the number of comparable measurements from other stations within the polygon;
μ is the number of comparable measurements within a polygon for the most numerous measurement type in the map area, determined before any stations have been eliminated.

In this way, scores maintain their balance with respect to the other criteria even though individual maps may vary considerably in the density and diversity of their measurement sets. If the calculation results in a value of more than 100 and m is greater than 1, S defaults to 100.

In order to avoid extreme results, the polygon area, A, has lower and upper default values. The lower default value is the area of the circle that encloses the space reserved for a station using the multiple-symbol format. The upper default value is 400 times the lower value:

$$A\min = \pi\left(R + \sqrt{2R^2}\right) / f$$

$$A\max = \left(\pi\left(R + \sqrt{2}R^2\right)^2 / f\right)400$$

where A is the area of a polygon;
R is the radius of the circular space allotted to a structural symbol, expressed in metres represented on the ground;
f is the number of square metres in the unit used to express polygon areas in the database.

The default values are calculated and displayed on the 'Output Selection' form, where they can be changed if desired.

The current scarcity of a measurement is what is important in determining its 'Scarce data' score and therefore the value of its station. For this reason, 'Scarce data' is the only one of the four criteria that requires recalculation every time a station is eliminated from the map.

Total Points

The total score for a measurement, T_{mnt}, is the sum of the points for all four criteria for that measurement, each multiplied by their percentage weighting factors, W, assigned on the 'Output Selection' form:

$$T_{mnt} = (K.W_K) + (C.W_C) + (V.W_V) + S.W_S)$$

The score for a station is the highest score for any one measurement at the station:

$$T_{stn} = T_{mnt}\max$$

In the single-symbol format, there is only one measurement at a station so the score for this measurement is automatically the score for the station. In the multiple-symbol format, only one of up to four measurements provides the score for the station. The other measurements, however, are not completely ignored because their influence is incorporated into the 'Complete dataset' score of the measurement that is used.

The point scores for stations are used to index the records in the 'Station IDs Points' table. Stations that potentially interfere with each other are then added to the map in sequence, those with the most points first. Any station that spatially conflicts with a station already on the map is omitted. It is important to note that stations are considered as a whole, even when plotting multiple symbols per station. Either all the selected symbols for a station are plotted on the map without overlap, or the station and all its symbols are omitted. This is consistent with the principle underlying the 'Complete data' criterion, that the importance of a measurement at a station depends in part on its relationship with other measurements at the same station.

All stations that have structural measurements score some points. Stations that consist only of an outcrop location score zero and are added in random order after the stations with measurements.

Classifying Stations

The process of deciding which stations interfere with which is iterative and time consuming, especially when the multiple-symbol format is chosen. Two steps are taken to make it more efficient.

First, all the stations in the 'Station IDs Points' table are copied to the 'Station IDs Test' table, where they are tested for possible interference. If a station is sufficiently distant from all the other stations for there to be no possibility of overlap, using the chosen plotting format, it is considered "free". Free stations generate records directly in the 'Final output' table and are deleted from the 'Station IDs Points' table. They require no further processing until final output begins.

Second, the remaining stations are assigned to clusters by a routine in the 'Station IDs Test' table. Each station in a cluster has a real or possible overlap with a neighbouring station assigned to the same cluster. Different clusters are separated from each other by areas of the map where there are no stations. These areas act as buffer zones that remove any possibility of mutual interference. The inclusion of a particular station on the map not only controls the inclusion of stations that directly overlap it, but it also controls stations that do not overlap it directly, but overlap a station that

does. The influence of any station on other stations is transmitted by a ripple effect from station to station to the edge of a cluster, but not beyond the cluster. Clustering divides stations up into groups that contain all the stations that need to be prioritized relative to each, and no others. This minimizes the number of calculations required to determine if a station should be added to the map or omitted. It is especially important in the process that optimizes the distribution of symbols in the multiple-symbol format.

Testing Stations

Once stations have been classified into clusters, the remaining records in the 'Station IDs Points' table are sorted first by cluster number and then by total point counts, and the 'Station IDs Test' table is cleared.

The station with the largest point count in the first cluster is then copied to the 'Station IDs Test' table, where it passes the overlap test by default because there are no other stations in the table with which it can overlap; at the same time it is deleted from the 'Station IDs Points' table. The station with the next largest point count in the cluster is then copied to the 'Station IDs Test' table and deleted from the 'Station IDs Points' table. In the 'Station IDs Test' table, it is tested to see if it overlaps the station already there. If it does, it is deleted from this table, but if not, it is retained.

Each station in the cluster is tested in similar fashion in order of decreasing point count. The last records to be tested are those that have no structural symbols and consist only of outcrop locations; these all have point counts of zero. If the multiple-symbol format is used, the assignment of symbols to sectors is recalculated for all stations presently in the 'Station IDs Test' table each time a new station is tested. This maximizes the chances that the new station will be able to fit without overlap.

Whenever a station is unable to fit without overlap, it is deleted from the cluster and will not appear on the map. It is also deleted from the subsidiary table where the 'Scarce data' criterion is calculated. This criterion is then recalculated, the total point counts for all the remaining stations are adjusted, and the stations within each cluster in the 'Station IDs Points' table are reordered as necessary.

When all the stations in a cluster have been tested, those that survived the testing procedure generate records in the 'Final output' table and the 'Station IDs Test' table is cleared for testing the next cluster. After all the clusters have been tested, the data are ready for output from the 'Final output' table.

Final Output

The various 'Station IDs' tables contain one record for each station. The 'Measurements' tables have one record for each structural measurement and may have several records for some stations and none for others. The 'Final output' table combines the data from both types of table, creating a record for every symbol, either structural or outcrop, that is to be plotted at each station (Figure 8). Thus, at least one record is present for each station, but stations may be represented by up to five records, depending on what data are available and what format has been chosen for plotting.

The simplest choice is to plot a single symbol on location. In this case, every station is represented by one record specifying that either a structural symbol or an outcrop symbol be plotted at the station's UTM coordinates. The most complicated situation arises under the multiple-symbol format. In this case, all stations are represented by a record specifying that an outcrop symbol be plotted at the station location. In addition, all stations having structural symbols have a record for each of these symbols specifying the nature and attitude of the symbol and the UTM coordinates of the plotting position, which is offset from the actual station location. A station having a full complement of symbols will have five records, one for the outcrop symbol and four for structural symbols, all with different UTM coordinates.

The 'Final output' table downloads the data in a form that can be read directly into the graphical component of the GIS for plotting. The output file provides the 'Station ID', the 'Symbol code' for each symbol, using the symbol scheme designated on the 'Output Selection' form, the 'Azimuth' and 'Dip/plunge' if applicable, and the UTM coordinates for the point where the symbol should be plotted. *GeoSymbol* has already adjusted all data for offsets and the scale of the map.

CONCLUSIONS

GeoSymbol is intended to complement *GeoLegend* in a GIS, the former controlling the structural and outcrop symbols while the latter manages the map units. Both are able to display the full detail of their datasets, if the scale of a map allows, but can generalize the data for less detailed scales. Consequently they make a digital geological map independent of scale. The two databases can be operated separately and are intended to work on different layers of a GIS. The only interdependence is in the need for *GeoSymbol* to extract certain polygon information from *GeoLegend* in order to use the 'Representative measurement' and 'Scarce

data' thinning criteria. Both databases could also be linked to the same 'References' table, as could most other geoscientific layers.

The present implementation of *GeoSymbol* is a prototype, which leaves much room for improvement. The present plotting formats use the simplest possible mathematical constructs based on circles. A more sophisticated method for outlining the spaces reserved for symbols would allow more compact plotting and more plottable stations on a map. The three basic plotting formats are more an example of what can be done than a comprehensive choice of all the useful options. Different situations call for different solutions, and this is especially so if *GeoSymbol* finds application with non-bedrock symbols, such as those that depict glacial striation data. In general, new formats can be created fairly easily by adapting the current ones.

The assumptions concerning symbol layout make programming easier, but sometimes at the expense of valid geological preferences. An example is the lack of support for special symbols indicating horizontal or vertical structures. Similarly, the specifications for symbol sizes may be too restrictive or it may be useful to allow various kinds of outcrop symbols. Many of these problems can be addressed within the present programming structure and resolving them would make digital maps significantly more user-friendly for geologists.

The criteria used for thinning stations have never been tried in real situations so their suitability is still uncertain. They attempt to mimic the decisions of a geologist by combining several issues into single numbers for each station, leading to a true or false result. The complexity of the problem prevents the method from being tested in any objective way. It requires that many maps be processed, using various permutations, and that these maps be subjectively evaluated to see if the results are acceptable. Some criteria may prove to be useless, others may need to be added, and the balance between criteria may need to be adjusted. Some of the criteria depend on arbitrary limits and default values, which may need revising. As in the case of the plotting formats, the criteria presently use relatively simple mathematics to arrive at their results and some refinement in this regard may improve precision.

Most importantly, *GeoSymbol* is presently too slow to be useful in an interactive environment. However, experience with *GeoLegend* has shown that professional programming, using more powerful programming tools can increase speed by several orders of magnitude, as well as making the database compatible with widely used computing platforms. It is intended that *GeoSymbol* will eventually be adapted to run within the data viewer being developed by the Geological Survey of Newfoundland and Labrador for distribution with its GIS datasets.

ACKNOWLEDGMENTS

I thank Gerry Kilfoil and Ken Matthews for mathematical advice. Larry Nolan provided insight into how computers process data, leading to increases in speed that make the database usable, and I have benefited from watching Andrea Bassan reprogram *GeoLegend*. Steve Ash plotted some experimental datasets to check that plotting formats and thinning work as intended. Gerry Kilfoil and Peter Davenport are thanked for reviewing the manuscript.

REFERENCES

Brodaric, B.
Undated: Fieldlog v2.83, tutorial/manual. Geological Survey of Canada.

Colman-Sadd, S.P.
1998: *GeoLegend*, version 1.0, for Helix Express: database and documentation. Newfoundland Department of Mines and Energy, Geological Survey, Open File NFLD/2650, version 1.0.

Colman-Sadd, S.P., Ash, J.S. and Nolan, L.W.
1997: *GeoLegend*: a database system for managing geological map units in a Geographic Information System. Computers and Geosciences, Volume 23, pages 715-724.

Phillips, F.C.
1960: The use of stereographic projection in structural geology. Edward Arnold (Publishers) Ltd., London, 2nd edition, 86 pages.