A Review of the MCWC CD-ROM Version 4.0:

Use Recommendations

By
Ralph J. Garono, Ph.D.
Earth Design Consultants, Inc.
800 NW Starker, Suite 31
Corvallis, OR 97330

Printed on Recycled Paper

TABLE OF CONTENTS

APPENDIX I

APPENDIX I

APPENDIX I

APPENDIX I

APPENDIX I

APPENDIX I

Introduction

The use of geographic information systems (GIS) in managing and analyzing environmental data in Oregon is becoming widespread. Although GIS technology has been available for 25 years or so (Marble 1998), recent technological advancements have made GIS available to many desktop computers users. Consequently, the availability of spatial data sets is also increasing. These data sets are often composed of information collected for different reasons by different groups at different times. Consumers of these GIS data layers include Oregon watershed groups, for example, the Tillamook Bay National Estuary Project (TBNEP), The Coos Bay Dynamic Estuary Management System Project (DEMIS) and now the MidCoast Watershed Council (MCWC), all of which have made GIS central to their watershed management projects.

Several watershed groups in Oregon have invested heavily in GIS by acquiring necessary computer hardware and software, training, and spatial data sets. The strategy to develop watershed management and monitoring plans seems to be to acquire readily available data sets and to use GIS as a tool to store, analyze, and communicate this information. However, like any other tool, GIS users must be aware of important limitations, especially limitations involving spatial data sets. Not all data are good data, and not all good data can be used all of the time: periodically, data must be critically reviewed for completeness and utility.

Development of a CD-ROM containing all readily available data represents an important first step in developing a MCWC GIS. Whether data are useful or not depends on the question being asked and on the quality of the data. Simply put, the availability of a GIS data layer does not guarantee that that data layer will be useful to a particular analysis. In fact, Marble (1998) cautions, as "GIS technology has become generally available, as well as more user-friendly, an erroneous notion has developed that these changes mean that he technology can be mastered by almost anyone with minimal effort." Because of the wide availability of spatial data sets and the assumptions and limitations that have gone into collecting and then mapping these data, there is an enormous potential for misuse.

The data in a GIS have three components: actual phenomenon or characteristics (value, name, etc.), its spatial location, and time (Dangermond 1990). Value results from measurement or observation that was undertaken to record a particular phenomenon. For example, a fish survey records the number of fish in a particular area. For the survey to be scientifically meaningful, the observer usually follows a protocol that dictates: the length of the period of observation or sampling effort, the locations to be sampled, how the data are summarized (average, standard deviation, etc.) and reduced (statistical comparisons), etc. Inherent in the measurement of the number of fish are the study design, sampling protocol, and the assumptions that go into data collection. These factors ultimately affect how representative a sample is of a particular location and how confident one can be in the results. Information that describes sampling protocols must somehow be bundled with the data.

In addition to biological survey data, other types of information can also be incorporated into spatial data sets. For example, data layers that describe land cover, road types, culverts, land ownership, precipitation, etc. are readily available. Techniques used to measure or observe these features include photography, road inspections, culvert surveys, examination of tax parcel information, or empirical relationships between elevation and rainfall (a model). The point is that in all of these cases, decisions are made by the builders of these data sets as to what information is included and what is not (e.g., whether a road is paved or not paved, or whether a particular field is shrubby or open). Assumptions and limitations, as well as the source data also need to be documented and bundled with the data layer.

Another component to Dangermond's notion of spatial data is location. Indeed, the map overlay is central to all GIS (see Chrisman 1990b). Just as there are assumptions made in recording measurements or observations during surveys or studies, there are also assumptions that go into map-making. Although a detailed discussion of map making is beyond the scope of this report, there are a few points to consider.

> First, the map and reality are not identical. Even the best map is an inaccurate representation of the real world. There is error in all maps. Chrisman (1990b) warns, “consumers of maps rarely understand the limits to map accuracy and more general issues of data quality”: in other words, never use a map without understanding how the map was biased by the methods used to make it.
>
> Second, maps can be metric or non-metric (informational, Maling 1989). For the most part, GIS layers are used to produce metric maps. Sometimes GIS layers are produced with the intention of showing the extent of some feature (information) and not necessarily intended for analysis. Examples of informational GIS layers may include historic vegetation or variable scale layers. Erroneous conclusions can result from using a non-metric map layer in an analytical situation.
>
> Third, "data and information are not synonymous" (Cracknell 1999, p. 493). Maps are produced using both raw data and/or weighted composites (value-laden). For example, juvenile salmon densities mapped to a GIS stream layer represent raw data, while maps of a “habitat suitability index” represent a weighted composite. Muehrcke (1986) warns, “care is needed when using maps based on value-laden composite measures.” It is not hard to see that if different indexes were created at different times by different groups with different responsibilities, then the index values might represent different information. There is “danger in the scientific appearance” of index maps (Muehrcke 1986).

> Finally, many different types of features can be mapped. Some features may appear to have angular and very precise boundaries, like parcel (cadastral objects) maps. Other features may have fuzzy boundaries or smooth curves, like natural features (e.g., lakes or wetlands: Chrisman 1990a). Parcel maps are not based on the perception of change between two components, as are natural features. Mapmakers decide how these features will appear on a map.

Metadata, data about data, describes acquisition methods, scale, sampling protocols, spatial error, mapping assumptions, etc. In order for a GIS data layer to be useful for watershed analysis, that layer must be accompanied by metadata. For example, without metadata, it would be nearly impossible to determine if year-to-year variations in fish densities (no. fish / m^2), stored in a GIS layer, were due to actual variations in fish densities, variability in the data (noise), differences in sampling methods, etc. Considering that the GIS layer containing fish abundance information may only be one of many GIS layers used in a watershed analysis, there are many opportunities for information to be misused. Indeed, error can creep into GIS layers from a variety of sources: experimental error, mapping error, digitization error, user error, just to name a few. The introduction of error into a GIS-based study is known as "error propagation" and is summarized in a report done by G. Benoit (1997), an Oregon State University, Marine Research Management student.

Purpose & Scope

This project was initiated to review the GIS layers present on the MCWC CD-ROM Ver 4.0. The CD-ROM contains all available spatial data and readily available metadata and represents a good first effort at an initial data inventory. The MCWC CD-ROM contains data sets similar to those collected by two other Oregon watershed projects, TBNEP and DEMIS.

The purpose of this report is to review and summarize the GIS data layers and associated data files. Information contained in this report and in Appendix I depicts the geographic extent of each data layer, scale, and how the layers were generated. This report also presents use restrictions (based on the experience gained in applying the data set to the Rock Creek Watershed and other projects) and provides recommendations.

Content of MCWC CD-ROM

The MCWC CD-ROM Ver. 4.0 is divided into 11 directories: 1) BIO, 2) ECON, 3) ENVQUAL, 4) GEOMORPH, 5) HYDRO, 6) INFRA, 7) LANDCOV, 8) ORTHOS, 9) POLIT, 10) SOFTWARE, and 11) TERRAIN. However, not all directories contain data. Directory number 7 (Ortho) is empty and directory 9 (software) contains GIS viewing software. Consequently, nine directories are included in this data review.

File types on this CD include GIS layers (Arc Info Coverages, ArcView Shapefiles, and Arc Info GRID files), database files (*.DBF), data files (*.DAT), and text files (*.TXT or *.HTML). Arc Info coverages and shapefiles can contain information as points, lines (arcs), or polygons. Arc Info GRID files contain information in a raster data matrix, or grid. Database files can be linked to Arc Info coverages so that the data contained in the database file can be represented as map coverages (i.e., spatially). The text files usually contain information about coverages, grids and databases (i.e., the metadata).

There were 164 data layers on the MCWC CD-ROM: 42 line coverages, 22 point coverages, 63 polygon coverages, 31 GRID files and 6 shape files.

Evaluation of Metadata Completeness

It is difficult to anticipate all of the uses of these data sets; therefore, a series of criteria were used to evaluate the level of documentation for each layer.

Criterion #1

Criterion #1 was to **determine if metadata were present on the CD-ROM**. The responses could be Y (yes) or N (No). In some cases, one general description of a layer was referenced to many layers that were derived from a single study or single data layer. In these cases, I marked the "Present on CD" with a N/R (Not present, but Related metadata available). *I recommend that each data layer be given its own metadata description because data layers could potentially become separated from the CD and used independently.* In several cases, the metadata consisted simply of a link to a web site (most often the State Service Center for GIS, SSCGIS) that contained the pertinent metadata. I scored these layers as 'No' metadata present on the CD. *I recommend that the metadata be bundled with the actual layer because one of the roles of GIS is to store data sets.* Data clearing houses, like the SSCGIS, frequently update layers and metadata; therefore, it is possible that a future user of the MCWC CD-ROM would no longer have access to the metadata that describes data on the CD.

Criterion #1: Less than half of the layers on the CD were associated with metadata. Forty-six layers were associated with metadata present on the CD and 70 were not. An additional 45 layers did not have metadata, but had pointers to metadata associated with closely related layers. Although these metadata files could be used to describe some of the aspects of the data layers, they were often missing important information. *I recommend that each layer have its own metadata file.*

Criterion #2

Criterion #2 was to **determine if the spatial error was documented** for each layer. This criterion pertains to the spatial nature of GIS data. In other words, how confident can one be of the location depicted on the map layer. A layer was scored 'Y' if the spatial error was explicitly stated in the metadata or if the scale of the data layer was known and uniform. Layers were also scored 'Y' if the metadata indicated that the layer did not conform to National Map Accuracy Standards (NMAS), provided that the scale was known. If the data layer was composed of variable or unknown scale data, then the layer was scored as 'N.'

Criterion #2: The spatial error was not documented for most layers. The spatial error was documented for 45 layers (or they used a documented base map). One hundred and sixteen layers did not have spatial error documented.

Criterion #3

Criterion #3 was to **determine if the methods used to create a layer were documented**. If the methods used to produce the GIS layer (i.e., identification of the base map or a description of the digitization process (on screen *vs*. from a paper map)) were described, then the layer was given 'Y.' If there was no mention of the methods used to create the layer, then this criterion was scored as 'N.'

Criterion #3: Documentation describing the creation of each data layer was missing for most layers. Forty- seven layers had documentation of the map layer construction (criterion#3) and 114 did not.

Criterion #4

Criterion #4 was to **determine if the methods used to collect the data (portrayed in the data layer) were documented**. This criterion pertains to the value (see discussion above) of the phenomenon that is being portrayed as a map layer. If the data collection methods were described, then this criterion was scored as 'Y.' If there was no mention of the methodology, this criterion was scored 'N.' *It is important to note that not all information portrayed as a GIS layer is the result of sophisticated studies. For example, property boundaries have very precise definitions compared to the boundary between a wetland and a field. No methodology is required in the case of the former, while there are strict protocols that are followed to define wetland boundaries.*

Criterion #4: Most layers did not have data collection methods documented. Thirty-six layers had the data methods documented, 90 did not (35 were not applicable).

Criterion #5

Finally, criterion #5 was to **determine if the fields in the data table were documented**. Many spatial data sets have both a map component and a data component. The data components are organized into a data table (similar to a spreadsheet) where the data records are link to map features. The column headings in the data table are the variable names. Metadata should generally describe what data are contained in each column of the data table. If the fields names (column headings) were described in the metadata, criterion #5 was scored as 'Y.' If there was no description of the data in the data table, this criterion was scored 'N.' A print out of the data table column headings is included for each data layer in Appendix I as 'field definitions.'

Criterion #5: Most layers did not have data fields names documented. Of the layers examined, 51 layers had the fields documented, 96 did not, and 14 were NA.

Detailed notes are given in **Appendix I** for each data layer reviewed. Data layers occur in this document in the same order that they appear on the MCWC CD. That is, they are organized by themes in the directories mentioned above. The following information is presented in a text box the data layer name, source, date of creation, type, projection, scale and CD directory. Answers to the five criteria described above also appear. A graphic shows the geographic extent of the layer. Below the text box is a description of how the layer was created. This description is composed of excerpts from the metadata, or related metadata, when available.

The second page shows the number of records appearing in the layer info table. This number corresponds to the number of data records attached to points, lines, polygons, or grid cells. This number is a reflection of how much information is attached to the data layer. This number is one factor that a researcher would look at to determine how useful a particular data layer would be to a study. For example, a culvert layer that has 12 records may be less valuable to a study than a culvert file with 1012 records. Following the number of data records is the file structure (field definitions). This table shows the column headings for the data file. Good metadata describes what the column headings are: some column headings are easy to figure out and others are more cryptic. Wherever possible, I mention additional information sources that are related to the layer being described. Finally, I present comments and use restriction/ recommendations.

Summary of Database Files

Database files also exist on the CD-ROM. These database files are reviewed in the table below and do not appear in Appendix I. Database files were reviewed from the 1) BIO, 2) ECON, and 3) ENVQUAL directories on the CD-ROM. Texts appearing in quotations are excerpts from metadata files.

1) BIO

Layer	Date	Metadata Present	Description
Coho_den.dbf	09/29/97	Y	"Relative coho densities are based on uncalibrated snorkel estimates. Differences between snorkelers, weather, and water conditions may all effect observations. Data should not be used to extrapolate population estimates." "Relative densities were estimated in pools only. Riffle data is not included."

Layer	Date	Metadata Present	Description
Comments	This database contains the average density of juvenile fish per pool for multiple years. Stream reaches are identified by name. To be useful for watershed analysis, averages should be located on segments of a hydrography GIS layer (1:24,000). In addition, averages themselves are simply measures of central tendency. These data would be more useful if the number of observations and some estimate variability were also included. Finally, some idea of how representative (i.e., repeatable) these samples are would also be an important consideration for those interested in using these numbers in analysis.		
juvsur96.dbf	1996	Y	"Stream reaches were selected because they were either standard coho spawning surveys, identified as important spawning and rearing for summer steelhead, as an additional site, or as part of a monitoring plan for central coast habitat restoration projects." "The juvenile surveys covered the entire length of the adult coho spawning survey stream reach, up to one mile or an easily identified ending point in other identified streams, or sampling every fourth pool habitat unit until ten units had been sampled in additional surveys. In all surveys every fourth pool habitat unit was snorkeled. Surveys were conducted during August and September. " Methods are well documented in *bio\juvrp96a.htm*. Due to the lack of resources, no calibrations of the snorkeler were conducted.

Layer	Date	Metadata Present	Description
Comments	This database contains the average density of juvenile fish per pool for 1996. It also contains information on vegetation and land use. Stream reaches are identified by name. These stream segments were not surveyed randomly; therefore, these data cannot be used to express the condition of all streams, in general. It is a biased sample. To be useful for watershed analysis, sample points should be located on segments of a hydrography GIS layer (1:24,000). In addition, averages themselves are simply measures of central tendency. These data would be more useful if the number of observations and some estimate variability were also included. Finally, some idea of how representative (i.e., repeatable) these samples are would also be an important consideration for those interested in using these numbers in analysis.		
mc_pass.dbf	unknown	Y	This database file contains information on fish passage. Locations are given as township, range, section and quarter quadrangle.
Comment	There is no date on this file. It would be helpful to know when these observations were made and how they were made (observation, measurements, etc.). It would also be useful to map these observations onto a 1:24,000 base map for future watershed analyses.		
salmon.dbf	unknown	N	This database file contains the ODF&W Statewide Salmon hatchery release data from 1980 to 1997 for 80 hatcheries.
Comments	There is no description of these data, the methods used, what the numbers represent, or estimates of error and variability. Methods and sample variability must be documented to be useful for analysis. In addition, it would be helpful to map the locations of these hatcheries on 1:24,000 base maps.		
spawn95.dbf	10/24/97	N	Adult Salmonid Spawning survey data (peak density and area under the curve) - links to 1:100,000 scale rivers coverages through dynamic segmentation. Data are from the early to mid 1990s. Species are given as numbers (there is no key).
Comments	There is no description of the methods used to collect these data, nor are there estimates of the data variability. Data are also linked to 1:100,000 scale base maps. Before these data are used, methods need to be documented and data variability assessed. If possible, transfer this information to larger scale (1:24,000) base maps for watershed analysis.		
stw-sts.dbf	unknown	N	ODF&W Statewide Steelhead hatchery release data from the 1980s and 1990s.
Comments	These data could be linked to hatchery locations.		

2) ENVQUAL

Layer	Date	Metadata Present	Description
wq_d*.dbf	unknown	Y	EPA Basins datafiles - Water quality (attach to coverage wqsites). Currently these datafiles hold statewide information and each correspond to a five year period. (*.aih, *.ain relate to these files).
Comments	Well documented water quality data that can be linked to 1:24,000 scale base maps for watershed analysis.		

3) HYDRO

Layer	Date	Metadata Present	Description
sil24-68.dbf sil69-93.dbf siu67-93.txt	Unknown	N	Daily flow values
Comments	Methods and units should be archived with these values. Numbers could be linked to 1:24,000 scale base map and used for watershed analysis.		

Map Scale

Map scale is an important characteristic of a spatial data set. Muehrcke (1986) points out that "Map scale is not independent of the dimensional character of the mapped phenomena." That is to say, that the level of detail generally increases as the scale of the map increases. For example, a river depicted on a 1:24,000 scale map shows many more meanders than the same river mapped at 1:100,000 scale. Moreover, a roads layer created at 1:24,000 scale often shows more detail than roads mapped at 1:100,000. It is easy to see that the results obtained from GIS analysis are dependent on the scale of the data used. Imagine calculating the density (miles per acre) of roads within a watershed using a detailed 1:24,000 roads layer *vs*. a 1:100,000 scale roads layer. If the larger scale layer (1:24,000) depicts more bends and twists (and perhaps even roads that don't appear on the smaller scale map), then the overall road length will be longer in the larger scale layer, than in the smaller scale layer. This will affect the calculated road density. The Oregon GWEB Manual suggests that watershed analyses be conducted at a 1:24,000 scale (Oregon Watershed Assessment of Aquatic Resources Manual, 1999, page I-8).

Positional accuracy is also dependent on map scale. For this report, if the scale of the map is known, and if the map adheres to National Mapping Accuracy Standards, then the spatial error is known. For example, if a map conforms to NMAS the horizontal positional accuracy for a discrete feature at 1:24,000 scale would be 1/50th of inch, or +/- 40 feet (12 m). At the same scale, NMAS requires that no more than 10 % of points tested can be more than ½ a contour interval off (Muehrcke 1986). (*It is interesting to note that a line 1/50th of an inch on a 1:100,000 scale map is the graphic equivalent of a corridor 167 feet wide!*)

Most data layers contained on the MCWC CD-ROM were at a map scale of 1:100,000 or they were unknown (Table 1). There were 26 layers at 1:24,000 or larger.

One important note is that 17 data layers were constructed from variable scale data. As mentioned, the scale at which the data layer is created is related to the level of detail and the positional accuracy of the mapped features. If the scale is variable, the data layer is not useful for analysis and can only be for "display purposes" (a phrase that occurs in the metadata).

Table 1. Summary of map scales of data layers on MCWC CD-ROM.

Map Scale		Layers on MCWC CD-ROM	
From	To	Number	Percentage
1:500,000	And up	6	3.7
1:250,000	1:500,000	1	0.6
1:100,000	1:250,000	44	27.3
1:50,000	1:100,000	3	1.8
1:24,000	1:50,000	0	0.0
1:12,000	1:24,000	26	16.1
	Variable	17	10.6
	GRID or Unknown	64	39.8
	Total	161	99.9

In summary, the map is really a model that represents real-world features. It is important to understand the limitations and assumptions that go into map making. Clarke (1990) warns that "digital cartographic data bases can be highly precise and many map users believe them to be accurate, when in fact, the data stored as digital maps are often unbelievably faithful reproductions of incorrect maps."

Recommendations for collecting new layers

1) New data layers should be developed at a 1:24,000 scale.
2) Document all methods for layer production, including digitization steps and accuracy assessments.
3) If data are to be transferred to another base map, go back to original data source.

4) Use well-documented base maps to depict information. For example, use standard DLG files (available from USGS) that have known spatial accuracy and adhere to National Map Accuracy Standards. Avoid creating layers from variable scale data.
5) For all summaries and layers created by analysis, document both the source layers and the steps taken to summarize layers.
6) For GIS layers that make use of data resulting from technical studies, document the study date, researcher, methods (protocol), how measurements were made, units of measurement, etc.
7) Bundle all metadata with GIS layers.

References & Suggested Readings

Benoit, Gregory M. 1997 An interactive metadata catalog for the Tillamook Bay National Estuary Project geographic information System. M. S. research report, Marine Resource Management, Oregon State University, Corvallis, OR.

Busse, K. Tillamook Bay National Estuary Project Watershed Assessment Modules, CD-ROM. October 1997. ***(Not Cited in TEXT, but may be a useful reference)***

Clarke, K. C. 1990. *Analytical and computer cartography*. Prentice Hall, New Jersey.

Chrisman, N. R. 1990a. Efficient digitizing through the combination of appropriate hardware and software for error detection and editing. In Peuquet, D. J. and D. F. Marble (eds.) Introductory readings in Geographic Information Systems. Taylor & Francis, NY.

Chrisman, N. R. 1990b. The accuracy of map overlays: a reassessment. In Peuquet, D. J. and D. F. Marble (eds.) Introductory readings in Geographic Information Systems. Taylor & Francis, NY.

Cracknell, A. P. 1999. Remote sensing techniques in estuaries and coastal zones – an update. Int. J. Remote Sensing. 19:485-496.

Dangermond, J. 1990. A classification of software components commonly used in geographical information systems. 1990. In Peuquet, D. J. and D. F. Marble (eds.) Introductory readings in Geographic Information Systems. Taylor & Francis, NY.

Coos Bay - Dynamic Estuary Management Information System CD-ROM April, 1998.

Maling, D. H. 1989. *Measurements from maps: principles and methods of cartometry*. Pergamon Press, NY.

Marble, D. F. 1998. Urgent need for GIS technical education: rebuilding the top of the pyramid. Arc News, Redlands, CA vol: 20 Spring, pages1, 28.

Monmonier, M. 1991. *How to lie with maps*. The University of Chicago Press, Chicago, IL. ***(Not Cited in TEXT, but may be a useful reference).***

Muehrcke, P. C. 1986. *Map use: reading, analysis and interpretation*. JP Publications, Madison, WI.

Oregon Watershed Assessment of Aquatic Resources Manual. 1999. Governor's Watershed Enhancement Board, Salem, OR.

Tomlin, C. D. 1990. *Geographic information systems and cartographic modeling*. Prentice Hall, NY. ***(Not Cited in TEXT, but may be a useful reference)***