Selected papers from the DCW data quality project

Sindre Langaas
Håvard Tveite

Project Report No. 1/1995

DCW & Data Quality

Selected papers from the DCW data quality project

Sindre Langaas
UNEP/GRID-Arendal

Håvard Tveite
Department of Surveying
Agricultural University of Norway

Project Report No. 1/1995

Project:
Issues of error, quality and integrity of digital geographical data: The case of the Digital Chart of the World

Funding sources:
Norwegian Research Council, Project No. 31783/223
UNEP/GRID-Arendal, Project No. 205026

Introduction

This project report is a compilation of three papers presented by involved scientists in the DCW Data Quality project at two occasions during the spring 1995.

The first event was a joint United Nations Environment Programme (UNEP)/Global Resource Information Database (GRID) & Consultative Group of International Agricultural Research (CGIAR) Workshop held in Arendal, Norway, 8 - 11 May 1995, with the aim *inter alia* to discuss mutual interests among CGIAR centres and UNEP/GRID centres in the field of GIS data. The first paper in this compilation was an invited expert presentation. The purpose of the presentation was to review current small scale cartographical databases that are of interest for several CGIAR and UNEP/GRID centres for strategic applications, i.e. applications covering large areas ranging from several countries to continental to global. The databases reviewed were the Digital Chart of the World (DCW), the World Vector Shoreline and a Digital Elevation Model currently under development and derived from the DCW. Also the issue of data quality was briefly touched upon.

The second event was the 5th Scandinavian Research Conference on Geographical Information Systems held in Trondheim, Norway, 12 - 14 June 1995. The second and third paper in this compilation both were presented at this conference. The second paper addresses the issue of accuracy assessments of geographical line data sets. The paper suggests an alternative measure to assess the positional accuracy of line data sets compared to current standards for spatial data, namely average displacement. The paper also proposes a new data quality component for linear features, shape fidelity, and a measure, oscillation, as an indicator of this. The calculation of both these measures involves the use of neighbourhood buffering iteratively followed by boolean overlay operations and generation of statistics. The third paper conceptually discusses the differences in the data quality component completeness as defined and applied in the data quality part of the US Spatial Data Transfer Standard and the data quality part of the European Standard currently under development by CEN TC287, respectively. The DCW is used to exemplify the differences.

List of papers

I. S. Langaas. 1995. Cartographical Data and Data Quality Issues. Presented at the UNEP/GRID and CGIAR Workshop (Arendal II), Arendal, Norway, 11-14 June 1995.

II. H. Tveite and S. Langaas. 1995. Accuracy Assessments of Geographical Line Data Sets: The Case of the Digital Chart of the World. In: J.T. Bjørke (ed.) Proceedings from the 5th Scandinavian Research Conference on Geographical Information Systems, 12-14 June 1995, Trondheim, Norway, pages 145 - 154.

III. S. Langaas and H. Tveite. 1995. To Characterise and Measure Completeness of Spatial Data: A Discussion Based on the Digital Chart of the World (DCW). In: J.T. Bjørke (ed.) Proceedings from the 5th Scandinavian Research Conference on Geographical Information Systems, 12-14 June 1995, Trondheim, Norway, pages 155 - 161.

Cartographical Data and Data Quality Issues

Sindre Langaas
UNEP/GRID-Arendal
c/o Dept. of Systems Ecology, Stockholm University, S-106 91 Stockholm, Sweden.
Phone: +46-8-161737 Fax: +46-8-158417 E-mail: langaas@grida.no

Abstract

This paper reviews three readily available cartographical databases (DBs), the Digital Chart of the World (DCW), the World Vector Shoreline (WVS) and 30 Arc-Second DCW Digital Elevation Models (DEM) all originating from the US Defense Mapping Agency (DMA). These are presumed strong candidates as cartographical data sources for strategic needs at several Consultative Group of International Agricultural Research (CGIAR) centres. While most cartographical themes found within these DBs are of acceptable quality, some themes definitively needs improvement. No known substitutes to these are known to the author. An approach is being suggested to overcome these inadequacies. Furthermore, the data quality part of the US Spatial Data Transfer Standard (SDTS) is being briefly summarised with the aim to stress the importance of data quality reporting within the UNEP/CGIAR project when data sets are created.

Introduction, definitions and scope

I will in this presentation briefly summarise information on some GIS databases (DBs) that originates from the US Defense Mapping Agency, are currently available, are potentially suited for strategic needs of several CGIAR centres and have been recommended to be used as such at the first CGIAR/UNEP workshop (**Arendal I**). *Strategic needs* was at Arendal I defined as requirements related to studies covering a large geographical area, for example in characterisation and classification to be used in strategic research planning (anon. 1992). I will limit this presentation to cartographical DBs. By *cartographical DB* is meant a structured collection of digital GIS datasets digitised from topographical (or equivalent) paper maps or created for the purpose of making topographic type paper maps. Commonly, the following themes are included in cartographical DBs: (i) Coastline, (ii) International boundaries, (iii) National administrative boundaries, (iv) Transport infrastructure, (v) Cities, (vi) Hypsography or Digital Elevation Models and (vii) Hydrography. In this presentation International and Sub-national Boundaries are excluded. These will be presented by Deichman and Fox later. Besides the value of this kind of GIS data for cartographical purposes, they are also of great value in GIS modelling. The use of administrative boundaries to derive statistics is a most prominent example.

I will not examine sources for GIS cartographical data for operational needs. *Operational needs* was defined at **Arendal I** to be related to much more detailed data requirements for studies applied to relatively small areas. The distinction between strategic and operational data are here set at a scale of 1:1,000,000 (or resolution 1 km^2).

Furthermore, I will briefly present the data quality part in the US Spatial Data Transfer Standard (SDTS). The SDTS is one out of several standards for geographical data (and information) that recently has been developed or are under development. The intent by presenting the data quality component of SDTS is to emphasise the need for this sort of meta-data information, increasingly important when GIS data are being combined in simple or complex models or being used for other purposes than the initial one(s). This presentation rely on experiences from several GRID-Arendal projects, Clark's (1992) presentation at Arendal I as well as information found in literature, analogue and digital on-line available from Internet

Cartographical data for strategic needs

Existing Datasets

Table *Global base layer data presented at Arendal I (from Clark 1992)*

	Database	**Size (MB)**	**'Ownership'**	**Scale**
1	Hershey	1.2	Public	> 1:40M
2	World Data Bank (WDB)-1	1.5	Public	> 1:12M
3	WDB-II	110	Public	> 1:3M
4	Micro-WDB-II	2.5	Public	> 1:10M
5	ARC/WORLD	?	Commercial	> 1:3M, 1:25M
6	Mundocarto	150	Commercial	> 1:1M
7	World Vector Shoreline (WVS)	150	Public	> 1:250K, 1:1M
8	Digital Chart of the World (DCW)	1,700	Public, Commercial	> 1:1M

In Arendal I the following cartographical DBs were described (Clark 1992): Please refer to Clark (1992) for descriptions of most of these DBs. I will here focus upon the Digital Chart of the World (DCW) and the World Vector Shoreline (WVS). These DBs were recommended at **Arendal I** as the most important ones to be used by CGIAR Centres at the strategic level (anon. 1992). I will also introduce a digital DEM currently being derived from the DCW. GRID-Arendal and GRID-Nairobi has applied the DCW in several projects and have therefore accumulated quite considerable experience with its potentials and limitations, both regarding data qualities and practical problems. In particular, GRID-Arendal is, together with the Dept. of Surveying, Agricultural University of Norway, carrying out a project aimed at examining data quality issues of the DCW (Langaas and Tveite 1994, 1995).

Digital Chart of the World

Geographical coverage, description of content, spatial resolution

The DCW was made from two map series, the Operational Navigation Charts (ONC, 1:1 mill.) and the Jet Navigational Charts (JNC, 1:2 mill. Antarctica only) by the US Defense Mapping Agency, and collaborating partners in the UK, Canada and Australia. (ESRI 1992). The DCW is a digital representation of the global ONCs and JNCs and therefore a cartographical DB. Its development is documented in DMA (1992a). Both map series are made with a large number of mapping rules reflecting their intended purposes. For the ONCs the purposes area given in the product specification (DMA 1981):

> *"The 1:1,000,000 scale Operational Navigation Charts (ONC) Program provides aeronautical charts to support medium altitude enroute navigation by dead reckoning visual pilotage, celestial, radar, and other electronic techniques. In the absence of Tactical Pilot Charts (TPC's), these charts should also satisfy the enroute visual/radar navigation requirements of pilots/navigators flying low altitude operations (500 feet to 2000 feet above ground level). The ONC is also used for operational planning, intelligence briefings, and preparation of visual cock-pit displays/ film strips essential to aerospace navigation of high-performance weapon systems."*

The thematic content of the map series and their digital representation, the 1.7 GB DCW, reflects these purposes. Hypsography, Drainage, Roads, Populated Places, Political/Oceans, Land Cover, Railroads, Utilities, Cultural Landmarks, Transportation Structure, Physiography and Aeronautical are the major themes. For a complete description of themes (layers) and features, please refer to, e.g., ESRI (1992) or http://sun1.cr.usgs.gov/glis/hyper/guide/dcw.

Custodian, availability and format

The DCW was first released on four CD-ROMs by DMA as public domain data in the Vector Product Format (VPF) for a cost of US$200. Actually, the main purpose of making the DCW DB was to promote the use of the VPF, a recently developed military GIS format (DMA 1992a). Following this release, the DCW DB has now been released by a large number of GIS software vendors in their own proprietarian formats. Most of these DCW DBs are being sold at different price levels. The DCW DB now exist in ARC/INFO, MapInfo, Atlas and Intergraph formats, besides the initial VPF format. There exist also a number of more and less robust public domain conversion tools ftp'able on Internet.

Practical access

For the DCW in VPF and ARC/INFO formats, the relevant information for obtaining the CD-ROMs can be found in the GRID-Arendal Directory of Environmental CD-ROMs. For the DCW in various commercial GIS vendor formats, information should be available at the local dealer. UNEP/GRID also have an agreement with ESRI that allows the various GRID-centres to provide extracts upon request.

Shortcomings in content and quality

Users of the DCW, including ourselves, have after some years of usage, identified a large number of deficiencies in the various themes. These deficiencies are mainly related to the initial purposes of the ONCs and JNCs and the specific mapping rules guiding the compilation of the map sheets and the time of compilation (and updating) of ONC maps. Many of the features in several of the themes should only be included in the ONCs when they were of navigational value according to the specifications. To give an example from DMA (1981) related to populated places:

> *"702. Density and Selection*
> *A. The following general rules are formulated to govern the selection of populated places.*
> *1. In areas where populated places are very numerous, a selection of cities, towns and villages shall be shown to a density commensurate with scale.*
> *2. In areas where populated places are generally sparse, cities, towns and villages shall be shown to a density comparable to the density on a standard 1:500,000 scale map of the area".*

Similar kind of mapping rules related to the navigational significance of the various features .are found for most themes

The age of the original maps is another importance quality factor. For example most of the African tiles dates back to the 60ies and 70ies. It should further be kept in mind that the map sheet production year may deviate from the source material age. Therefore, several of the themes are not of a sufficiently high quality to be used for modelling purposes. Those that have been found by GRID-Nairobi to hold a quite high level in Africa are Hypsography, Drainage, Populated Places, and Political/Ocean boundaries (Goff 1994). Those found to be so-called problem coverages were the Roads, Railroads and Utilities. The rest where found somewhere in-between.

Processing requirements / problems

Due to the considerable data amounts, quite powerful HW and GIS SW are highly recommended. From own experiences we would strongly recommend UNIX type HW and SW, at least during DCW data preparation and editing phases. Further, it has been experienced that some themes, such as the Drainage layer, have had number of arcs per tile exceeding the software limitations of, e.g., UNIX Arc/Info. Although possible to bypass

Maintenance and update problems

The DMA is currently working on the second edition of DCW. We believe that this version will only remove errors related to transfer from paper maps to digital data, such as coding errors, duplication of lines, mislabelled edges, etc. and to a much lesser extent deviations from reality. Therefore, the actual content will predominantly remain the same. It is therefore quite obvious that many CGIAR centres will need to edit, update and correct DCW data. One technical solution to this can be to scan at high resolution another paper map at higher accuracy of the region of interest, to reproject the raster map and DCW to the same projection system and then to edit the DCW vector data superimposed upon the scanned (raster) map. Edit here means to fit to the positioning of the same geographical feature in the other map source.

World Vector Shoreline

Geographical coverage, description of content, spatial resolution

The World Vector Shoreline is a global dataset of shorelines created by the US Defense Mapping Agency (IOC et al.1994). It was developed by the DMA at a nominal scale of 1:250,000. Global coverage was complete in 1989. The sole feature is the shoreline. The primary data source for the WVS was DMA's Digital Landmass Blanking (DLMB). These were deduced primarily from the Joint Operations Graphics and coastal nautical charts also produced by the DMA. The DLMB data is a raster data set with 3 by 3 arc-second interval geographic grid, which explains the 3 arc-second stepping interval in the WVS when displayed at large scales. For parts of the world not covered by the DLMB, the shoreline was taken from the best available hardcopy sources at a preferred scale of 1:250,000.

Custodian, availability, practical access and format

The WVS is in the custody of the US DMA. A beta test version of the WVS was released on CD-ROM in the VPF format also used for the DCW described earlier. Another version of the WVS was released in 1994 by the International Oceanographic Commission (of UNESCO) and the International Hydrographical Organisation as part of the General Bathymetric Chart of the Oceans (GEBCO) on CD-ROM (IOC et al.1994). The latter version is stored in an internal binary format. Luckily, as a part of the GEBCO CD-ROM there a exist a conversion utility that enables conversion to DXF format, edible or importable by most GIS software.

Positional accuracy

The specification for positional accuracy, is that 90% of all identifiable shoreline features should be located within 500 meters (i.e. 2 mm at 1:250,000) of their true geographic position with respect to the World Geodetic System (WGS-84) datum. The precision as defined by the 3 by 3 arc-second steps imply at worst around 100 m at Equator and improving towards the Poles.

Comparison with digital coastline data from Norway in the same scale (1:250,000) made by the Norwegian Mapping Authority does not reveal differences of any significance worth mentioning (J.-A. Bordal, pers. comm.). Although far from most CGIAR centres' regions of interest, we tend to believe that the accuracy is good for most parts of the world. For strategic needs the WVS is of high quality.

30 Arc-Second DCW Digital Elevation Models (DEM)

Geographical coverage, description of content, spatial resolution

Terrain and hypsography information represented by digital contour lines or by Digital Elevation Models are becoming increasingly important for both analytical and visualisation purposes. Height information is provided in the DCW as contour lines at selected and unequally spaced intervals (ESRI 1992). For many purposes an equally spaced (in x and y direction) DEM is more suited. Therefore, USGS represented by the EROS Data Center, in association with UNEP/GRID-Sioux Falls, are currently developing a consistent 30 arc-second global DEM. The 30 arc-second resolution of course varies with latitude. At Equator this equals a raster cell size of approx. 930 m.

The width decreases with latitude to the half at 60° N and S. The method used to create the DEM applies both the contour and point height information as well as the drainage network information from the DCW. As of 15. January 1994, Africa, Haiti, Madagascar and Japan were ready, and the South-America in advanced progress. For more up-to-date information about the global progress the Customer Services EROS Center or UNEP/GRID-Sioux Falls can be contacted.

Custodian, availability, practical access and format

The above mentioned already prepared DEMs can be obtained from the EROS Data Center by at least two ways, through Internet by ftp (for free) or on unlabeled CCTs (for a modest cost). It was also planned to be released on CD-ROM. The DEM data are provided as 16-bit straight raster (also termed unsigned 2-byte binary data) images. The height information is provided as feets. 4 ancillary files provide additional meta-information on issues such as file structure, world co-ordinates and position information. One of these files supports the ARC/INFO Image Integraton routine for image-to-world transformation.

Data quality

The absolute accuracy of the vector information in the DCW is 2000 meters circular error (horizontal) and ± 650 meters linear error (vertical) at 90 percent confidence according to the specifications (DMA 1992b). The DEM grid created obviously will be no more accurate than its sources. We are not aware of assessments that have been carried out to evaluate the accuracy of any of the sub-sets of this global DEM under preparation.

DATA GAPS

When one compares the cartographical themes offered by the three above-described DBs with those normally included by cartographical DBs, the situation looks quite good. The World Vector Shoreline is an obvious candidate for coastline data due to superior resolution and accuracy compared to the DCW. The DCW still has much to offer cartographically. However, as clearly illustrated in the initial aim of the ONC and JNC map series and by practical experiences, there are several themes that require improvement in data quality, in particular positional accuracy and completeness (see below) before fulfilling the needs of the CGIAR Centres and others. Transport infrastructure and Hydrography are two cartographical themes that seriously needs improvement. We are not aware of new data source at strategic scales that will resolve this need in the near future. An approach based on the existence of scanned higher quality paper maps and subsequent on-screen editing of co-registered DCW data is suggested as one feasible way of editing.

GIS data quality issues

With the quite revolutionary developments taking place in the field of digital geographical information technology and use the last decade, the need for meta-information to accompany the digital GIS data sets have become apparent. In particular, the possibilities to use GIS data sets for multiple purposes and several data sets to be combined in multi-layer model have enforced this. Several standardisation efforts have been initiated. One example is the US Spatial Data Transfer Standard developed by the joint US geodata community under the leadership of USGS (Fegeas *et al.* 1992). We here use the SDTS as an example, firstly because it is one standard already ready and implemented (NIST 1992), secondly because it is accessible over Internet.

While standards such as the SDTS encompass many parts, we will here specifically address the data quality part. We consider this of major importance within the joint UNEP/CGIAR project as a main aim of the project is to jointly compile, distribute and maintain high quality natural resource and socio-economic digital data sets. The data quality report of the SDTS consists of five portions being:

- lineage
- positional accuracy
- attribute accuracy
- logical consistency
- completeness

We here will briefly review these five parts, with the objective that the CGIAR centres compiling GIS data sets attempt to cover these quality parts in their data reporting. The following brief summary is taken from NIST (1992).

Lineage

This part shall include a description of the source material, methods of derivation, including all transformations involved. Appropriate dates should be included for relevant data sources and processing steps.

Positional Accuracy

This part shall include the degree of compliance to the spatial registration standard (another part of SDTS). Quality of control assessments shall be reported by using the procedures established in the geodetic standard. Descriptions of positional accuracy shall consider the quality of the final product after all transformations. The date of any positional test shall be included. Variations in positional accuracy shall be reported either as additional attributes of each spatial object or through a quality overlay (reliability diagram). Four optional methods for measuring positional accuracy are suggested.

Attribute Accuracy

Accuracy measurements for attributes on a continuous scale shall be performed using procedures similar to those used for positional accuracy. The report of a test of attribute accuracy shall include the date of the test and the dates of the materials used. In the case of different dates, actual changes in the phenomena shall be described.

Spatial variations in attribute accuracy may be reported in a quality overlay. Three quantitative methods are suggested for attribute accuracy assessment

Logical Consistency

A report of logical consistency shall describe the fidelity of relationships encoded in the data structure of the digital spatial data. A number of tests can be carried out to assess various types of logical consistency, such as -
- Do the data contain permissible values only ?
- Do lines intersect only where intended ?
- Are any polygons too small, or any lines too close ?

The term "topologically clean" is allowed to be reported provided that
(a) All chains (arcs) intersects at nodes. Use of exact tolerance shall be reported.
(b) Cycles of chains and nodes are consistent around polygons. Or, alternatively, cycles of chains and polygons are consistent around nodes.
(c) Inner rings embed consistently in enclosing polygons.
The quality report shall report software (name and version) and dates of tests.

Completeness

Completeness in SDTS refers to information of selections criteria, definitions used and other relevant mapping rules such as minimum area or minimum width. Deviations from standard coding schemes, as well as definitions and interpretation shall also be reported.

Conclusions

The World Vector Shoreline and the Digital Chart of the World complemented by a DEM derived from the latter are believed to provide the CGIAR centres with most necessary cartographical GIS data at the strategic level (< 1:1,000,000). Some themes definitively need improvements. Which (DCW) themes and how much improvement is needed will depend upon the various centre needs in terms of data quality. To stress the significance of and need for proper geodata quality reporting, the quality part of the US Spatial Data Transfer Standard is briefly reviewed.

Acknowledgements

This paper has benefited considerably from the DCW Data Quality Project funded by the Geographic Information Technology Programme of the Norwegian Research Council and UNEP/GRID-Arendal. Critical comments by Christopher Smith, GRID-Arendal, improved this paper considerably.

References

Clark, D.M. 1992. Basic data layers for resource analysis and evaluation. Discussion paper prepared for the CGIAR/NORAGRIC/UNEP Meeting on Digital Data requirements

Defense Mapping Agency. 1981 - . Product specifications for Operational Navigation Charts (Code: ONC) Scale 1:1,000,000. First edition 1981 and changes and amendments thereto. Defense Mapping Agency, Washington., D.C.

Defense Mapping Agency. 1992a. Development of the Digital Chart of the World: Washington, D.C., U.S. Government Printing Office.

Defense Mapping Agency. 1992b. Military specification Digital Chart of the World (DCW). MIL-D-89009. 204 pages.

ESRI. 1992. The Digital Chart of the World for use with ARC/INFO® Data Dictionary. ESRI, Redlands, CA.

Fegeas, R.G., Cascio, J.L. and Lazar, R.A. 1992. An overview of FIPS 173, The Spatial Data Transfer Standard. *Cartography and Geographic Information Systems*, Vol. 19, No. 5., ??-??.

Goff, E. 1994. Evaluation of Digital Chart of the World for Africa. Internal Note, GRID-Nairobi.

IOC, IHO and BODC. 1994. Supporting volume to the GEBCO Digital Atlas. Published on behalf of the International Oceanographic Commission (of UNESCO) and the International Hydrographical Organisation as part of the General Bathymetric Chart of the Oceans (GEBCO); British Oceanographic Data Centre, Birkenhead, UK.

Langaas, S. and H. Tveite. 1994. Project proposal: Issues of error, quality and integrity of digital geographic data: The case of the Digital Chart of the World. URL http://ilm425.nlh.no/gis/dcw/dcw.html

Langaas, S. and H. Tveite. 1995. To Characterise and Measure Completeness of Spatial Data: A Discussion Based on the Digital Chart of the World (DCW). Paper to be presented at ScanGIS'95, the Fifth Scandinavian Research Conference on GIS, Trondheim, Norway, June 12th - 14th, 1995.

National Institute of Standards and Technology. 1992. Federal Information Processing Standard Publication 173 (Spatial Data Transfer Standard). US Dept. of Commerce, 199 pages.

Accuracy Assessments of Geographical Line Data Sets, the Case of the Digital Chart of the World[*]

Håvard Tveite

Department of Surveying
Agricultural University of Norway, P.O.Box 5034, 1432 Ås, Norway
fax: +47 64948856 phone: +47 64948840
email: lanht@nlh.no

Sindre Langaas

UNEP/GRID-Arendal
c/o Dept. of Systems Ecology, Stockholm University, S-10691 Stockholm, Sweden
fax: +46 8 158417 phone: +46 8 161737
email: langaas@grida.no

Abstract

To be able to utilise geographical data for analysis, one should know something about the quality of the data. In present geographical data standardisation proposals (SDTS, CEN TC287), several aspects of geographical data quality have been described, such as lineage (data collection and processing history), spatial accuracy, attribute accuracy, completeness, logical consistency and currency.

Methods for quantitative assessments of different aspects of spatial accuracy for data sets of linear geographical features, such as shape fidelity and positional accuracy are described. For these assessments, independent data sets of better (and preferably known) accuracy will have to be used. In order to be able to do automatic assessments, data set completeness must be taken into consideration.

The method is to be used for assessing the spatial accuracy for some themes of the Digital Chart of the World (DCW) (scale of original maps (ONCs): 1:1000000), using the Norwegian mapping authority's national N250 map series (scale 1:250000)[**] as a reference data set.

Key words: Accuracy, geographical, digital, data, line, buffer, overlay, DCW

Introduction

The availability of quality information is a prerequisite for the utilisation of geographical data sets.

Traditional geographical maps have conveyed quality information indirectly through the quality constraints and mapping rules that applies to the relevant map series and implicitly through the (presentation) scale of the maps. The professional map users have hopefully been aware of the many aspects of traditional map quality, while most casual map users probably have used the scale of the map as the only quality indicator.

With the advent of digital geographical information, presentation scale as such is no longer a useful measure of geographical data quality since digital geographical information can,

[*] This work has been partially funded by the Norwegian Research Council under the Geographical Information Technology programme.

[**] Many thanks to the Norwegian mapping authority for giving us access to excerpts of the digital N250-data set for these purposes.

theoretically, be presented at any scale. The availability of digital geographical data and geographical information systems (GIS) also gives new opportunities for easy combination geographical data sets of any scale. The results of analysis on combinations of data sets depend on the quality of all the participating data sets.

In order to be able to determine the quality of the results of geographical data analysis, it is imperative that quality measures are available for all the participating data sets.

The inclusion of quality measures for digital geographical data sets has been impeded by the lack of standards. There has been some research activity on spatial data quality, and we some significant contributions include: Chrisman 1984, Goodchild and Gopal 1991 (book of articles), SDTS 1990 (US spatial data transfer standard).

The research presented in this article is a part of the ongoing project[*] «Issues of Error, Quality, and Integrity of Digital Geographical Data: The Case of the Digital Chart of the World (DCW)» (Langaas and Tveite 1994). Until now, we have been investigating methods for quality assessments, and are now starting to apply the methods using our data sets (DCW and N250).

The rest of the paper is structured as follows. In chapter 2, linear geographical phenomena are introduced. In chapter 3, different ways of measuring geographical line quality are discussed, and our method for quantitative assessment of geographical line quality on the basis of data of higher geometric accuracy is presented. Chapter 4 is a discussion of scale as it applies to geographical lines. Chapter 5 rounds it up with conclusions and an outline of future work.

Linear geographical phenomena

The geometric line abstraction can be used to represent many geographical phenomena. Some examples:

- Roads and railways
- Administrative (state, municipality) and economical (property) borders
- Utility lines (powerlines, telephone lines, water and sewage tubes)
- Rivers and streams
- Natural boundaries (e.g. vegetation, soil)
- Shorelines

Some of these phenomena are human «constructions» and some are nature given (and of course, most human constructions are constrained by nature).

There are many ways of providing quality measures for such linear features. The choice of a quality measure depends to some extent on the type of linear feature we are considering.

«Scale» and fractal behaviour

The «scale» of a line data set can to a certain extent be determined on the basis of the geometry of the line alone. Geometric accuracy is in many cases closely related to «scale». Good indications on scale are:

- The number of significant digits in the representation of point in the data set is the crudest measure of «scale» / spatial accuracy of a data set. This is not a useful measure

[*] The project presently has a WWW page: URL:http://ilm425.nlh.no/gis/dcw/dcw.html

when the original data have been manipulated (e.g. transformed to a new projection), as most software do not consider accuracy in their calculations.

- Distance between neighbouring points. The intended scale of the data set can normally be derived from the lowest distance between neighbouring points. This is not true if the data set has been manipulated, for instance by inserting new points on the lines using some sort of interpolation method.

- Frequency of curvature change. For curving phenomena which change curvature at a higher frequency than can be captured using the assumed geometric accuracy in the data set of interest, the maximum rate of curvature change is a good indication of the «scale» of the data set. Such phenomena are phenomena that show fractal behaviour (Barnsley 1988) up to larger scales than what can be expected by the data set under consideration. Most features in nature seem to exhibit fractal behaviour over a large spectrum of scales. Examples of such phenomena are: rivers/streams, roads, shorelines and other natural boundaries. The fractal behaviour of natural phenomena, and to a certain extent also human-made linear objects, is influenced by the soil/geology/geomorphology of the area.

Fractal behaviour of infrastructure

When one gets to a large enough scale, infrastructure will cease to exhibit fractal behaviour. A road will normally not change curvature more frequently than each 100 meter (1000 meters for a modern motorway, while perhaps 10-20 meters for a small older road). The same applies to railways, powerlines, telephone lines and other utilities. When you come to a certain point, infrastructure will cease to exhibit fractal behaviour. The fractal behaviour of infrastructure is, in addition to cultural/historical issues, also influenced by the geomorphology of the area.

Methods for assessing the quality of lines

In the following sections, we will be presenting and discussing methods for calculating and quantifying the geometric accuracy of lines.

For our assessments, we assume that we have two independent data sets, X and Q, covering the same line theme and the same area (and collected at about the same point in time). One of the data sets, Q, should have a known geometric accuracy. The geometric accuracy of Q should be at least an order of magnitude better than the expected geometric accuracy of the data set X. It is also expected that the completeness and consistency of data set Q is significantly better than that of data set X.

Lines

The geometric accuracy of a line can be decomposed into two components:

- Positional point accuracy: Positional accuracy can easily be given for well defined points on the line (e.g. the end-points). For the rest of the line, it is difficult to say anything about positional accuracy and to quantify it.

- Shape fidelity: To be able to say something about the accuracy of a line, it is useful to talk about its shape fidelity as compared to another line. The shape fidelity should indicate to what extent the curvature of two lines are similar.

The type of spatial «errors» that can occur for linear data sets could also be classified into categories. E.g.:

- Scale-dependent errors (generalisation). These are errors that result from reducing the sampling frequency when collecting data on the linear phenomena of interest.

  - Generalisation/sampling: A line-representation that has been generated by sampling a line of high geometric accuracy represents a special case. Each point of the line is very accurately specified, but between the represented points, there can be large deviations between the interpolated line and the original position of the linear feature. This is closely related to scale-dependent errors.

- Achievable accuracy of fuzzy lines. The position of most linear phenomena get fuzzy as the scale gets larger, and it is generally impossible to give them an *exact* location. River centrelines and soil and vegetation boundaries are good examples of fuzzy natural phenomena, but also human constructions can be difficult to measure with extremely high accuracy (it is difficult to determine the centreline of a road with millimetre accuracy).

- «Random» errors. Errors that result from erroneous sampling and data processing.

It would be desirable to be able to separate these when describing the spatial accuracy of the geometric representations of linear geographical features.

Point measures

It is straightforward to calculate the geometric accuracy of points. For single points one can measure the deviation vector (**e**) of the point representation (**P**) as compared to another representation of the same point with better (and known) geometric accuracy (**Q**).

$$\mathbf{e} = \mathbf{P} - \mathbf{Q} \qquad = (P_x\text{-}Q_x,\ P_y\text{-}Q_y,\ P_z\text{-}Q_z) \qquad \text{for 3D space}$$

The absolute value of this deviation vector ($|\mathbf{e}| = \sqrt{\mathbf{e}_x^2 + \mathbf{e}_y^2 + \mathbf{e}_z^2}$ for 3D space) is a useful measure for further (standard) statistical calculations.

For multiple points one has to resort to statistical measures to determine quality parameters. Standard deviation or variance can be used whenever the point-errors of the data sets have no bias and can be considered normally distributed.

The mean error vector (spatial bias) is:

$$\text{mean(e)} = \text{mean(P-Q)} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{P_i} - \mathbf{Q_i})$$

In the case of no point error bias (|mean(**e**)| = 0), the variance and standard deviation of the point errors (|**e**|) are:

$$\text{var}(|\text{e}|) = \frac{1}{N}\sum_{i\in(1..N)}|\mathbf{e}_i|^2$$

$$\text{SD}(|\mathbf{e}|) = +\sqrt{\frac{1}{N}\sum_{i\in(1..N)}|\mathbf{e}_i|^2} \quad (= \mathbf{E_{RMS}})$$

Both of these measures are acceptable quantifications of the spatial accuracy of points.

End-points

Line end-points can be used to provide a simplified measure of the geometric accuracy of the lines. End-points could be cross-roads and dead ends in a road network, river meets and lakes in a river/watercourse system or joints and end-points in a tube network.

If one is able to identify corresponding end-points in the reference data set and the data set of unknown spatial accuracy, it will be straightforward to compute a statistical measure of the geometric accuracy of the end-points using the formulas presented above.

Previous work on quantitative quality assessment on the DCW was performed using 40 evenly distributed cross-roads in the road and railroad network in the area covered by ONC G18 (the south-west coast of USA.), and using 1:100000 scale topographical data (US DLG) as reference data sets (1:24000 data were used for testing vertical accuracy). This work is described in a DMA report (DMA 1990).

Intermediate points

As long as intermediate points are not well-defined features, the only way of finding corresponding intermediate points is to search for the closest point on the other line. A method for determining spatial accuracy of a line as compared to a line of better accuracy could then be to traverse the line, and at regular intervals (spacing **e**) along the line take out sample points, and on the basis of each of these points do a search for the closest point on the reference line. At each sample point, the distance vector, **e**, to the closest point on the reference line is an indication of the spatial accuracy of the line at that point, and an overall measure of line accuracy can be calculated statistically using **e** as in the formulas presented above.

This method should be applied for all lines that have corresponding lines in the reference data set, arriving at an overall measure of the positional accuracy of the lines in the data set.

The choice of spacing **e** could be based on the spatial accuracy of the reference data set. Since the lines we are interested in does not exhibit completely random behaviour, this implies that the smaller **e** that is chosen, the more strongly will the **e**'s of neighbouring point samples be correlated. To get an overall statistical measure for the data set, **e** should therefore be chosen so large that the **e**'s of neighbouring points can be considered not correlated (Cov($\mathbf{e}_i$, $\mathbf{e}_{i+1}) \approx 0$). **e** could be chosen to be of a higher order of magnitude than the accuracy of the reference data set. It could also be interesting to do several calculation based on different **e**'s to give an assessment of the stability of the calculated spatial accuracy.

To determine separate measures for the line end-points and the interior of the lines, a transformation will have to be performed on each individual line prior to the traversal of the line, in such a way that the end-points of the corresponding lines match exactly.

Calculating the geometric accuracy of lines using buffering

The method proposed below uses buffering of lines and subsequent overlay analysis to give a quantitative assessment of the geometric accuracy of a line relative to another line (of higher accuracy). The method should be iterative, because it will not be possible to determine an optimal buffersize in advance (we do not yet know the spatial accuracy of the line data set under consideration). The size of the first buffer can be determined on the basis of the known spatial accuracy of the reference data (e.g. the standard deviation, SD, if that is available). For each iteration, the size of the buffer could then be doubled. 4-5 iteration will probably be sufficient, and the process should be terminated when the results seem to stabilise.

Before starting the iterative process it is useful to do some statistical calculations on the lines. The interesting measure at this point in the process is the total length of each line

The iterative process:

For each buffersize bs_i:

$bs_i, i \in \{1,2,3,\ldots,n\}$ (bs_i is the width of the buffer)

perform the following 3 steps:

First step - line buffering

Perform a buffer operation on each of the two lines, X and Q, using the buffer size bs_i (resulting in a buffer 2 x bs_i wide). Call the resulting polygons for Xbs_i and Qbs_i.

Second step - overlay

Perform an overlay of the two polygons Xbs_i and Qbs_i, the result being a new polygon data set: XQbs_i.

Third step - statistics

Calculate statistics (total area, number of polygons, total perimeter, perimeter/area for each polygon) on XQbs_i for the following situations:

- areas inside Xbs_i but outside Qbs_i (A($Xbs_i \cap \overline{Qbs_i}$))
- areas outside Xbs_i and inside Qbs_i (A($\overline{Xbs_i} \cap Qbs_i$))
- areas inside Xbs_i and inside Qbs_i (A($Xbs_i \cap Qbs_i$))
- areas outside Xbs_i and outside Q bs_i (A($\overline{Xbs_i} \cap \overline{Qbs_i}$))

Arriving at a measure for the geometric accuracy of lines

The statistics calculated in the above steps can be used to give measures of deviation of the line X from the line Q.

Average displacement

$$DE = bs_i \cdot \frac{A\left(Xbs_i \cap \overline{Qbs_i}\right)}{A\left(Xbs_i\right)}$$

DE is the lower bound of the average displacement of a line relative to another line (of greater accuracy in our case).

Oscillation

$$O = \frac{\# A\left(Xbs_i \cap \overline{Qbs_i}\right)}{Length(X)}$$

Where #A(...) is the count of areas.

O is an indication of the oscillation of the lines X and Q relative to one another.

This measure is most useful for «randomly» oscillating phenomena, where it could be used as an indication of bias (there would probably be a bias if the oscillation, O, is low for randomly oscillating lines of different accuracy).

Oscillation could also be found directly using X and Q, by counting the number of nodes introduced when overlaying the two line data sets.

O is also a measure of relative scale for «randomly» (that is random appearance at the relevant scales) oscillating linear phenomena.

Calculating the geometric accuracy of line data sets

The buffering method for calculating the geometric accuracy of lines can also be applied to line data sets. To apply the method on the data set level, all lines must exist in both data sets (the completeness criterion). If there are lines that only are present in one of the data sets, these will introduce errors in the calculations. In conjunction with spatial accuracy assessments on linear data sets, it is important that an assessment is made of the relative completeness of the data sets.

Calculating completeness for line data sets using buffering

Using an approximate measure of geometric accuracy of a data set (X), it is possible to make an assessment of the completeness / number of miscodings of the X data set, as compared to the Q data set. An approximate measure of the geometric accuracy can be obtained by applying the method presented above one the complete data sets (ignoring the lack of completeness measures).

The method outlined below use a combination of buffering, overlay and selection (and thinning).

First step - buffer

Perform buffering on both line data sets, X and Q, using a buffer distance, BD, which could be about twice as large as the geometric accuracy measure found for data set X (for the line-polygon alternative presented below, a buffersize of four times as large as the geometric accuracy measure found for data set X should be used to obtain the same statistical effect).

It is necessary to choose the buffer distance larger than the statistical measure of the spatial accuracy (could be SD), since SD is a sort of weighted mean. When choosing a buffer distance twice as large as the SD for both line data sets, we capture all errors within 4SD's of the reference line.

The result of this buffering is the data sets XB and QB.

Second step - overlay

Do two line-polygon overlays: Overlay X with QB and XB with Q, resulting in the new mixed data sets XQB and XBQ.

Third step - statistics

Using XBQ, calculate the sum of the length of the lines outside XB and compare it to the total length of lines in Q:

Completeness(X) = $100 \cdot \left(1 - \frac{length(\overline{XB}Q)}{length(Q)} \right)\%$

A more «exact» measure can be obtained by using the identity of the lines that are not in X, and calculate the length of the complete lines, as opposed to the part of the lines that do not fall within the buffer.

Using XQB, calculate the sum of the length of the lines outside QB and compare it to the total length of lines in X. This is a measure of the amount of miscodings in X as compared to Q:

Miscodings(X) = $100 \cdot \left(1 - \frac{length(X\overline{QB})}{length(X)} \right)\%$

This can also be done more «exactly» in the same way as described above.

Ensuring completeness

To prepare for the spatial accuracy assessment to come, all miscoded lines in X and all lines in Q that are not in X should be removed from the line data sets. The lines to be removed can be found in XBQ and XQB, described above. The resulting data sets should be used in the rest of the process.

Assessment of the spatial accuracy of line data sets

The process for calculating geometric accuracy of line data sets is exactly the same as for individual lines. It is, however, useful to start out with calculating the total length of the lines in both coverages.

The (iterative) process is exactly as described for single lines above:

1. Line buffering
2. Overlay
3. Statistics

Arriving at a measure for the geometric accuracy of line data sets

The statistics calculated in the above steps can be used to give measures of the deviation between the lines of the X and the Q data set.

A lower bound on average displacement for complete line data sets

$$DE = bs_i \cdot \frac{A\left(Xbs_i \cap \overline{Qbs_i}\right)}{A\left(Xbs_i\right)}$$

DE is a lower bound on the average displacement of a quality line data set relative to a line data set of less accuracy. The choice of reference data set will influence DE. We have chosen to use the data set with the smallest expected total line length as reference.

If the data sets operated on is the original data sets, as opposed to the completeness adjusted data sets, the results must be corrected using the completeness measures determined above, giving an approximate lower bound on average displacement for incomplete line data sets.

$$DE = bs_i \cdot \frac{A\left(Xbs_i \cap \overline{Qbs_i}\right) - (1 - Completeness(X)) \cdot Qbs_i}{A\left(Xbs_i\right) \cdot (1 - Misconding(X))}$$

Oscillation

$$O = \frac{\# A\left(Xbs_i \cap \overline{Qbs_i}\right)}{Length(X)}$$

Where #A(...) is the count of areas.

This is an indication of the oscillation of the lines X and Q relative to one another.

O is most useful for «randomly» oscillating phenomena, where it could be used as an indication of bias (there would probably be a bias if the oscillation, O, is low for randomly oscillating lines of different accuracy).

Oscillation could also be found directly using X and Q, by counting the number of nodes introduced when overlaying the two line data sets.

What’s next?

In this paper we have outlined a method for quantitatively assessing the spatial accuracy of the representation of geographical linear features. The method utilises the standard GIS operations buffer and overlay to arrive at a polygon data set that can be analysed using simple statistical measures (e.g. sum and count).

At the time of this writing, we are about to start our accuracy analysis of the DCW data set using these methods. The results of these practical exercises will become available to the public in the project report.

References

Barnsley, M., 1988, *Fractals everywhere* (Academic Press).

Chrisman, N., 1984, The Role of Quality Information in the Long-Term Functioning of a Geographic Information System. Cartographica, vol. 21, no. 2/3, pp. 79-87.

DMA, 1990, Digital Chart of the World - DCW Error Analysis. Prepared by Environmental Systems Research Institute, Inc, USA for Defense Mapping Agency, USA.

Goodchild, M., and Gopal, S., 1991, *Accuracy of Spatial Databases* (Taylor &Francis).

Langaas, S. and Tveite, H., 1994, Project Proposal: Issues of Error, Quality, and Integrity of Digital Geographical Data: The Case of the Digital Chart of the World. URL: «file://ilm425.nlh.no/pub/gis/dcw/quality.ps».

SDTS, 1990, Spatial Data Transfer Standard, version 12/90. USGS.

To Characterise and Measure Completeness of Spatial Data: A Discussion Based on the Digital Chart of the World (DCW)

Sindre Langaas

UNEP/GRID-Arendal,
c/o Dept. of Systems Ecology, Stockholm University, S-106 91 Stockholm, Sweden.
Fax: +46-8-158417 Phone: +46-8-161737 E-mail: langaas@grida.no

Håvard Tveite

Department of Surveying,
Agricultural University of Norway, P.O.Box 5034, N-1432 Ås, Norway.
Fax: +47-64948856 Phone: +47-64948857 E-mail: lanht@nlh.no

Abstract

There is an increasing degree of sophistication associated with describing the qualities of spatial data. Completeness is one data quality component that is included in both the US Spatial Data Transfer Standard (SDTS) and the European standard under development within the framework of the European Committee for Standardisation (CEN). While both standards use the same term, there are apparent semantic differences reflected in their definitions and proposed ways for assessment and reporting. This paper will discuss these differences and their implications using the global 1:1 million scale Digital Chart of the World (DCW) database as a test case.

Keywords: Completeness, SDTS, DCW, spatial data quality

1. Spatial data quality characterisation and measures

The change from paper maps to GIS **data** in various kinds of geographical data analysis and applications has made it easy to use the same spatial data for different applications and also for combing several layers into quite complex spatial models. This has created a need for data quality descriptions and measures to be attached to the datasets (whatever definition used of dataset). Thereby the user can judge the suitability for an intended application. A commonly used definition of quality is 'fitness for use' (Chrisman 1984). Further, if several spatial datasets with appropriate quality measures are combined, the error propagation can be modelled. Here Veregin's hierarchical error model comes to mind (Veregin 1989).

In Figure 1 is shown a modified version of Veregin's (1989) well-known model of the hierarchy of needs in modelling of error in GIS operations. We have adapted this model to more recent terminology and found that a hierarchy of needs for handling of spatial data quality better reflects the current status in terminology. In this model, level 1 is concerned with classification and identification of spatial data qualities. The efforts dedicated to the classification of spatial data qualities are reflected in the data quality parts of several on-going standardisation efforts. Level 2 focus on the characterisation and assessment of the qualities defined in level 1.

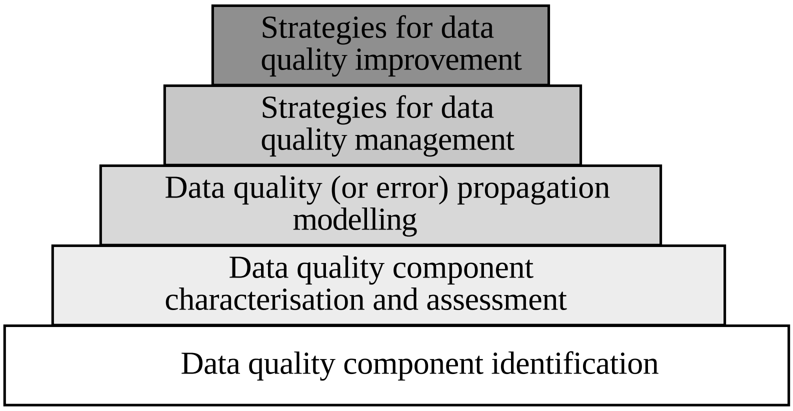
Strategies for data quality improvement
Strategies for data quality management
Data quality (or error) propagation modelling
Data quality component characterisation and assessment
Data quality component identification

Figure 1. A hierarchy of needs for handling of spatial data quality. The text and concepts in this figure, based upon Veregin (1989), is modified according to recent terminology

Several countries or groups of countries within both the civilian and military sector of the spatial data community have for a number of years worked on standards to facilitate transfer and use of spatial data (Moellering 1992). Two major efforts are the US Spatial Data Transfer Standard (SDTS, Fegeas *et al.* 1992) and the European standard currently under the development of European Committee for Standardisation (CEN) Technical Committee 287 (CEN/TC287/WG02 1995). These standards include data quality components.

SDTS including Part 1 with the data quality report specifications was approved in July 1992 and is currently being implemented by federal, state and private spatial data producers in the USA. The European Geographic Information standard with its Data Description - Quality part is currently being developed and is supposed to be ready by 1997/98. Both standards have, within the data quality part of their specifications singled out a quality component termed *completeness***.** While both standards use the same term, there are apparent semantic differences reflected in their definitions and proposed methods for characterisation and assessments.

In this essay we will briefly describe and discuss some of these differences. We will do so in view of experiences from an on-going project aimed at reporting of data quality of the Digital Chart of the World (DCW, ESRI 1992, Langaas and Tveite 1994). We want to highlight some aspects relevant to the usefulness of the two different completeness concepts and their suggested reporting characteristics and measures.

2. Completeness - reporting characteristics, measures and metrics

2.1 SDTS

The term *completeness* is not defined explicitly in the SDTS. It is stated, though, under the completeness section that 'the quality report shall include information about selection criteria, definitions used and other relevant mapping rules.' Further, 'The report shall describe the relationship between the objects represented and the abstract universe of all such objects. In particular, the report shall describe the exhaustiveness of a set of features. Exhaustiveness concerns spatial and taxonomic (attribute) properties, both of which can be tested.' The concept 'abstract universe of all such objects' is a key concept which in each case needs an accurate definition (or specification) to give the necessary information about the various completeness aspects.

In these specifications of completeness characteristics it appears that a cartographical digital database (or dataset) rather than a geographical digital database has been in mind. The distinction between cartographical and geographical databases is visualised in Figure 2. Here it is seen that 'reality' for cartographical databases are modelled twice. First, to create maps using not only strict objective thematic criteria but also cartographical criteria for readability and aesthetic purposes, and secondly these map(s) are modelled to derive a digital database. In a conventional thematical map production process there exist a wide range of selection criteria, specific definitions and other mapping rules that convey information about the suitability of the digitised version also for other purposes than the initial thematic one.

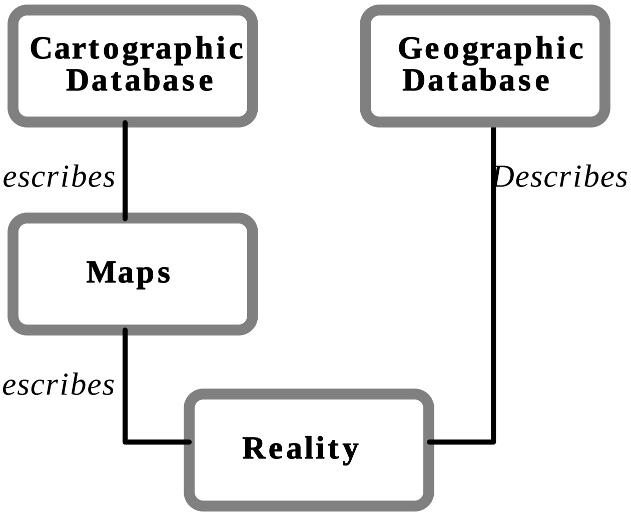

Figure 2. While geographical databases model and describe reality directly, cartographical databases do this indirectly.

In the SDTS, completeness reporting is primarily supposed to be done as textual reports and to a lesser extent as quantitative measurements, although it is referred to objective tests that can be carried out.

2.2 CEN/TC287/WG02

In CEN/TC287/WG02 (1995) completeness is defined as 'the difference between an actual dataset and its specifications.' It is further stated that 'completeness measures indicate how well the information reflects the content defined by the specification'. Taking this quantitative approach, three possible measures are suggested to quantify completeness. These are *omission*, *commission* and *coverage ratio***,** represented by the following metrics:

- Percentage of data missing relative to specification,
- Percentage of data present that is not in current specification of dataset or extract, and
- Occurrences of one variable per unit of another.

The CEN/TC287/WG02 assessment approach is more concise compared to SDTS. However, given its definition and recommended approaches of assessments, being solely quantitative, the precise definition of 'dataset' (what is a dataset ?) and 'specifications' given in quantitative terms are crucial for implementation. Furthermore, while the recommended approaches appear well suited for geographical datasets (or databases), they are more difficult to implement for cartographical datasets.

3. Completeness reporting of the DCW - some considerations

3.1 DCW - a cartographical database

The completeness quality aspect is of particular relevance for DCW. The digital DCW was made from two map series, the Operational Navigation Charts (ONC, 1:1 mill.) and the Jet Navigational Charts (JNC, 1:2 mill. Antarctica only). The DCW is a digital representation of the ONCs and JNCs and therefore a cartographical database. Both map series are made with a large number of mapping rules reflecting their intended purposes (DMA 1981):

“The 1:1,000,000 scale Operational Navigation Charts (ONC) Program provides aeronautical charts to support medium altitude enroute navigation by dead reckoning visual pilotage, celestial, radar, and other electronic techniques. In the absence of Tactical Pilot Charts (TPC’s), these charts should also satisfy the enroute visual/radar navigation requirements of pilots/navigators flying low altitude operations (500 feet to 2000 feet above ground level). The ONC is also used for operational planning, intelligence briefings, and preparation of visual cock-pit displays/ film strips essential to aerospace navigation of high-performance weapon systems.”

3.2 Completeness - the issue of ‘ideal’ reporting level exemplified

The quality reporting ideally should be assigned to various levels of the dataset. SDTS distinguishes between the following levels:

- Dataset (or database)
- Theme
- Map (or geographical extract)
- Feature/object (or thematical extract)
- Element

To clarify the difference between these levels, an example will be given.

An environmental researcher would like to use the DCW to quantify potential annual increase in methane (CH_4) releases from *cranberry bogs* in Northern Finland under doubled atmospheric CO_2 levels and associated temperature rise. *Cranberry bog* is one class or feature under the layer (or theme) Land Cover in the DCW. Completeness descriptions on the entire dataset level might be of limited relevance. However, the knowledge on the specific purposes of the ONC and JNC map, (i) aerial navigation and (ii) military strategic planning, obviously indicates that the information contained on cranberry bogs might be unsatisfactory. The next level of reporting is the theme level. Cranberry bog constitutes one class or features out of many in the theme (or layer) Land Cover of the DCW. A completeness description on the theme level, supposedly valid for the spatial extent of the entire datasets, then will provide more detailed information about the suitability for annual methane emissions. The next level of reporting might be the feature/object level. If specific completeness information is available on the cranberry bogs *per se*, then the researcher would be even better prepared to evaluate the suitability of the DCW for its planned application. Although not so relevant in this case, one might also find that completeness reporting down to the element level can be provided. Depending upon the spatial coverage of the dataset in question, the completeness reporting on the four levels; (i) dataset, (ii) theme, (iii) feature/object and (iv) element ideally should be provided for specific regions. Individual map sheets in the ONC or JNC charts are an obvious sub-division of the entire dataset region into smaller specific regions for reporting. Evidently, completeness reporting on cranberry bogs on the feature/object level for those map sheets that cover Northern Finland would be the most specific and useful completeness reporting that could be provided.

3.3 Quantitative completeness assessments of DCW

The SDTS recommends a topological test as the only quantitative (or structured) approach for completeness assessments besides textual reporting. In the European standardisation efforts, the quantitative approaches are the only ones recommended. Could these be employed for the DCW ? At the database level - hardly. 1.5 GB of digital data effectively prohibits this. At the theme level - hardly, but more feasible if reference data is available. When coming down to the geographical and thematical extract levels this becomes, theoretically at least, more attractive. Completeness assessment using the suggested measures *omission, commission* and *coverage ratio* requires (i) precise and quantitative specifications and (ii) relevant reference data that are presumed to be of a higher quality. Higher quality in this context means that the reference data comply better with the specifications given for the various themes and geographical regions. Such reference datasets are virtually non-existent given the purpose of the original map series referred to earlier and the associated detailed mapping specifications described in DMA (1981). One can, however, apply other existing and more general purpose geographical datasets that thematically are quite similar to the themes of the DCW. From a user perspective different from the initial ONC and JNC purpose - mirrored in the DCW database - this is quite satisfactory. Most users of the DCW are not using it for the aerial navigational and military planning purpose. Therefore, for the example given in para. 3.3 quantitative assessment with the recommended measures omission and commission for cranberry bogs in Northern Finland, provided that such digital data of high quality exist are feasible, would be highly attractive to the environmental researcher. However, this assessment would not give 'the difference between an actual dataset and its specifications.' The specifications as given in DMA (1981) is:

> “Rice fields, cranberry bogs and “similar flooded areas” shall only be shown when they are very unique or distinctive features in areas devoid of landmark detail.”

It is obvious that this specification is very subjective and renders testing almost impossible .

3.4 DCW completeness reporting - what do we do ?

Within our DCW Data Quality project we have chosen the proposed SDTS completeness understanding and approach for assessment reporting. This is more directed towards cartographical databases than the completeness part of the data quality section of the European standard under development. In practice, this means to summarise and structure the definitons and specifications given in DMA (1981). It should be mentioned though that Lineage and Usage part of the European standard does allow for extensive textual information. The completeness information or actual cartographical mapping rules instead can be reported in these parts.

Acknowledgements

This work has been partially funded by the Norwegian Research Council under the Geographical Information Technology Programme (Project No. 31783/ 223) and UNEP/GRID-Arendal (Project No. 205026)

References

Aalders, H. J.G.L., A. Giordano, and O. Jacobi. 1995. Quality: Discussion paper for an ESF GISDATA conference on quality. Published by the GISDATA Secretariat, Dept. of Town & Regional Planning, University of Sheffield, UK.

CEN/TC287/WG02. 1995. Geographic Information - Data Description - Quality. Draft for discussion produced by project team 5 (PT05), 1995-1-24, CEN Central Secretariat, Brussels, 36 pages.

Chrisman, N. 1984. The role of quality information in the long-term functioning of a Geographic Information System. In: Proceeding of AUTOCARTO 6, Vol. 2, ASPRS, Falls Church 1983, pp. 303-321.

DMA. 1981 - . Product specifications for Operational Navigation Charts (Code: ONC) Scale 1:1,000,000. First edition 1981 and changes and amendments thereto. Defence Mapping Agency, Washington., D.C.

ESRI. 1992. The Digital Chart of the World for use with ARC/INFO® Data Dictionary. ESRI, Redlands, CA.

Fegeas, R.G., J.L. Cascio, and R.A. Lazar. 1992. An overview of FIPS 173, The Spatial Data Transfer Standard. *Cartography and Geographic Information Systems* **19**(5): ??-??.

Langaas, S. and H. Tveite. 1994. Project proposal: Issues of error, quality and integrity of digital geographic data: The case of the Digital Chart of the World. URL file://ilm425.nlh.no/pub/gis/dcw/quality.ps

Moellering, H. 1992. *Spatial data base transfer standards: Current international status.* International Cartographic Association Commission on data standards, Elsevier Applied Sciences, NY.

Tveite, H. and S. Langaas. 1995. Accuracy assessments of geographical line datasets, the case of the Digital Chart of the World. In this Volume.

Veregin, H. 1989. Chapter 1. *Error modelling for the map overlay operation.* In: M. Goodchild and S. Gopal (eds.), Accuracy of spatial databases, Taylor & Francis, pages 3-18.