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 preferrably 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 has been applied 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

1 Introduction

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

[*] 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.

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 in theory can be presented at any scale. The availability of digital geographical data and geographical information systems (GIS) also gives new opportunities for easy combination of 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 involved 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 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 on our data sets (DCW and N250).

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

2 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

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

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

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

2.1 «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 points in the data set is the crudest measure of «scale» / spatial accuracy of a data set. This is not a useful measure 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 often influenced by the soil/geology/geomorphology of the area.

2.1.1 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, they 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.

3 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 better (preferrably at least an order of magnitude) 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.

3.1 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**).

$\mathbf{e} = \mathbf{P} - \mathbf{Q}$ $= (P_x\text{-}Q_x, P_y\text{-}Q_y, P_z\text{-}Q_z)$ 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 (P_i and Q_i are corresponding points in the two data sets):

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

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.

3.1.1 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).

3.1.2 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 **ε**) 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 has to 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 **ε** could be based on the spatial accuracy of the reference data set. Since the lines we are interested in do not exhibit completely random behaviour, this implies that the smaller **ε** 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, **ε** should therefore be chosen so large that the **e**'s of neighbouring points can be considered not correlated (Cov($\mathbf{e}_i$, $\mathbf{e}_{i+1}$) ≈ 0). **ε** 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 calculations based on different **ε**'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.

3.2 Calculating the geometric accuracy of a line 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 increased. 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 the lines.

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

3.2.2 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.

3.3 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 real linear data sets, it is therefore important that an assessment of the relative completeness of the data sets is made and used as a correction in the method.

3.3.1 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 on 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 that is 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.

This can also be done in a more «exact» way using in the same method as described above.

3.3.2 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.

3.3.3 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 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

3.3.4 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.

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

3.3.5 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.

4 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.