

Assessment of spatial quality - implications on calculation of area

By

Hans Skov-Petersen (hsp@kvl.dk)
Danish Centre for Forest, Landscape and Planning
The Royal Veterinary and Agricultural University
Hoersholm Kongevej 11
DK-2970 Hoersholm
+45 35 28 18 16

Date: 01/09/2004

Abstract: Error exists only between reality and measurements. Assessing the difference between alternative versions of measurements - e.g. by digitizing - can give important information about the methods used for measurement. Based on this the presentation introduces a number of classic and new approaches to assessment of inaccuracy as implemented in ArcGIS. Finally it is demonstrated how inaccuracy of digitized lines influences the accuracy in calculated areas by means of Monte Carlo simulation.

1. Introduction and scope

At a symposium on Uncertainty in Geographical Information in Helsinki, Finland in June 2003 (Virrantaus 2004), Professor Michael Goodchild from University of California, Santa Barbara told a story. And so it goes: Somewhere on the Globe, Professor Goodchild had been asked to produce a cadastral map of a plain, suburban dwelling area. The map was to be based on manual digitizing of paper maps in 1:1000. The customer wanted each parcel to be annotated with the cadastre number and the area in m^2 with – and this is the crucial point – 2 decimals. Professor Goodchild started wondering how feasible it would be to indicate the area with such a level of accuracy.

Now, one rule of thumb frequently used is that planar spatial accuracy of tablet-digitisation is ± 0.5 mm. So, somehow it is fair when looking at a standard parcel of 20×30 m (= $600 m^2$) to set the upper and lower limits of possible calculated area to an inner receptively outer buffer of e.g. 0.5 m (= 0.5 mm. in scale 1:1000). See figure 1.

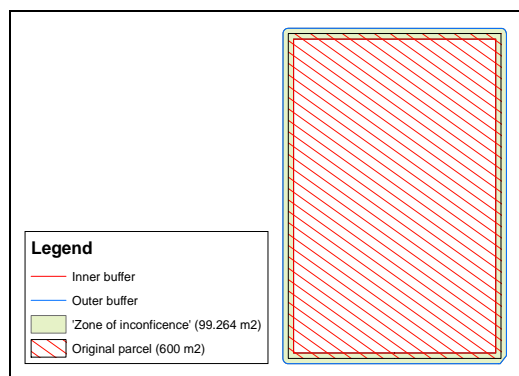


Figure 1: A 'Classical' approach to assessment of inaccuracy in area calculation

This means that, if the digitizing inaccuracy of ± 0.5 mm. holds some truth, the magnitude of the inconfidence is 99.264 m^2 of the area measured (see table 1)! In other words at best the measure can only be expected to be credible with two or three digits *before* the comma!

Table 1: Results from a ‘Classical’ approach to assessment of inaccuracy in area calculation

Text	Area
Original parcel	600 m ²
Inner buffer (0.5 m)	551.000 m ²
Outer buffer (0.5 m)	650.264 m ²
‘Zone of inconfidence’ (outer – inner buffer)	99.264 m ²

Putting it another way (see table 2), the map required to obtain an area calculation with two significant decimals ($\pm 0.01 \text{ m}^2$), should have an approximate scale of 3:1 in order to show the right magnitude of inaccuracy, i.e. the map should be three times larger than reality!

Table 2: Area-inaccuracy at different scales, based on a 20x30m parcel and a digitizing inaccuracy of ± 0.5 mm.

Scale	Inaccuracy (m digitizing scale)	Outer area (m ²)	Inner area (m ²)	Inconfidence (m ²)
1:1000	0.5	651.00	551.00	100.0000
1:200	0.1	610.04	590.04	20.0000
1:40	0.02	602.00	598.00	4.0000
1:8	0.004	600.40	599.60	0.8000
1:2	0.0008	600.08	599.92	0.1600
3:1	0.00016	600.02	599.98	0.0320
16:1	0.000032	600.00	600.00	0.0064
78:1	0.0000064	600.00	600.00	0.0013

Now we have (at least) two different stories: Firstly, *very much care should be taken when selecting the number of decimal digits* when using areas for annotation. Secondly – since the results appear to be surprisingly harsh on the expectations on magnitude of the accuracy of area calculations (but fully in order with the findings of Professor Goodchild) - *the method used for assessment of inaccuracy in area calculation might be taken to its extreme.*

Expecting digitizing-accuracy to be in the order of ± 0.5 mm. is a well established rule of thumb, but the anticipation that all vertexes will move in the worst possible direction simultaneously to gain the ‘inner’ respectively ‘outer’ buffers used above is highly unlikely to happen.

To provide a better assessment of the area calculation – and thereby a better idea of the number of significant digits to be used for annotations – two things are needed: A better model of the digitizing-inaccuracy and a new method for applying the inaccuracy-model to the calculation of areas. The remainder of this paper is providing a set of pragmatic attempts to fulfil these needs.

2. Method

2.1. Overview

As an alternative to the ‘classical’ buffer-method outlined in the previous paragraph a new method for evaluation of impact of digitizing-inaccuracy on measures of polygon-area is suggested. The basic idea is to relinquish the very strict interpretation of the ± 0.5 mm. inaccuracy used in the classical method. It is assumed that the digitizing errors are normal-distributed, individually for each vertex of the polygons. If such an inaccuracy model (mean and standard deviation of the coordinates (assuming normal distribution) could be established, a new copy of the original polygon - including the inaccuracy - could be simulated. If a (vast) number of different copies could produced this way as a Monte Carlo Process, a model of the inaccuracy of the polygon area could be established.

To evaluate the two techniques – the classical and the Monte Carlo - an experiment was set up. The test was performed in 1:2.000, 1:10.000 and 1:25.000. A single block of fields of 3 ha was selected randomly (see figure 2). Estimates of levels of inaccuracy of the areas were obtained for the three scales used. To obtain the inaccuracy model four copies were digitized for each of the three scales. Based on the inaccuracy model, 50 versions of the original block, for each of the scales, were produced. The mean and standard deviation of the areas calculated for each of the 50 versions was used as inaccuracy model of the areas. A flowchart of the process is found in figure 3.



Figure 2: The field block (approximately 3 ha) – highlighted just below the middle of the picture - used for the present test.

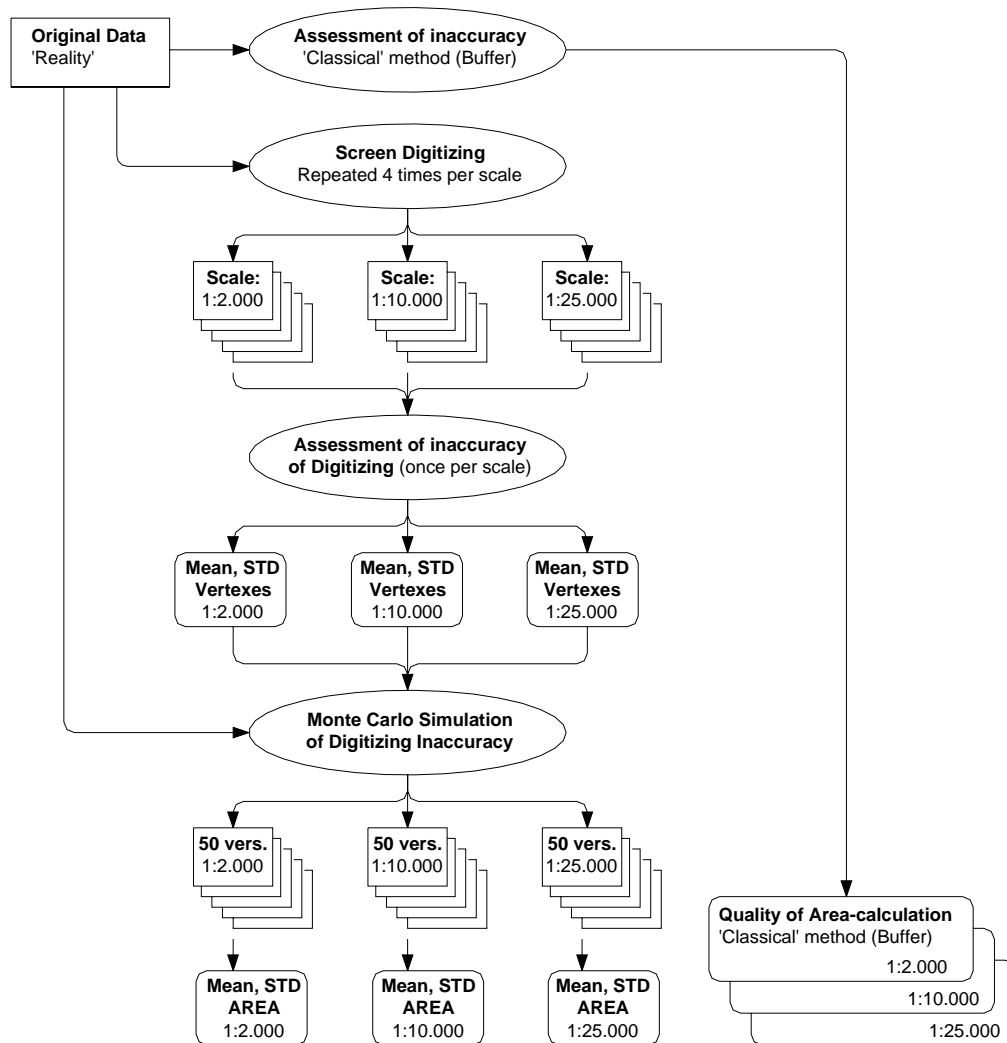


Figure 3: Flowchart of the process.

2.2. Application of the 'classical' method of assessment of area-calculation

For each scale inner and outer buffers corresponding to ± 0.5 mm. were constructed.

Table 3: Buffer distances used.

Scale	0.5 mm
1:2.000	1 m
1:10.000	5 m
1:25.000	12.5 m

2.3. Construction of an inaccuracy model of the digitizing process

To establish a model of the inaccuracy of the digitizing process a number of versions were compared with the original polygon (believed to be 'reality')¹. The comparison was done by

¹ It is on purpose that I throughout the paper refrain from using the term 'error'. In my view, an error is the deviation between a measurement and reality. Since a measure cannot be compared to reality – only with measurements of it – measures of 'error' seems to be an artefact that cannot be obtained. 'Inaccuracy' is the collected fuzziness of the data-capture and -handling processes. It comprises measurement techniques, semantics, scope of use etc.

measuring the orthogonal distance from each vertex of the version to be tested and the arcs of the ‘original’ (see figure 3). For each version base-statistics were calculated – including min, max, mean, std, fractiles etc. (see figure 4). The method was (in the first place) implemented as a generic AML-programme². The programme was based on the NEAR-Command. The programme calculates the statistics for mutual deviations for any combination of an infinite list of coverages or shape-files. Global statistics for all combinations in one is also provided.

Based on this method, standard deviations of inaccuracy for the digitizing process for each of the three scales could be obtained.

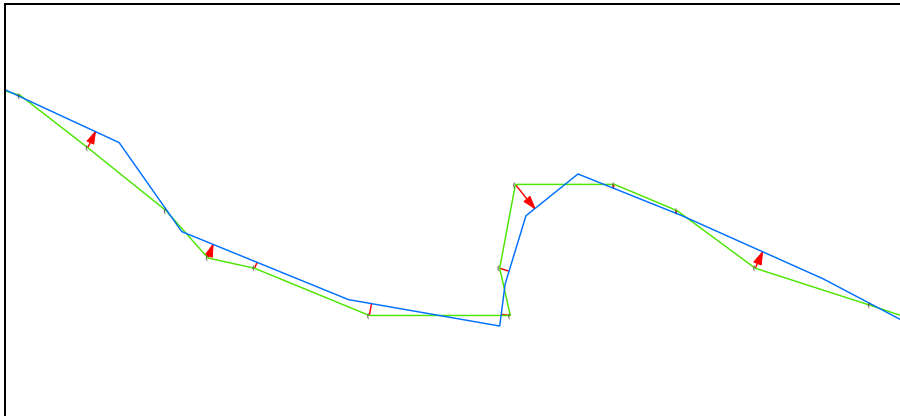


Figure 3: Illustration of the basis of the assessment of the differences between two versions of a data. Test-version in green, ‘original’ in blue and distance-vectors as red arrows.

Rowid	IN	OUT	MAX	MEAN	MIN	STD	Q25	Q50	Q75	LENGTH	LEN_PCT	COUNT	CNT_PC
1	arc2	arc3	0,032678	0,011396	0	0,009991	0,002824	0,009030	0,017281	2,643023	101,454806	1	100
2	arc2	arc4	0,133708	0,036007	0,000000	0,031976	0,010083	0,027329	0,055638	2,643023	94,286060	1	100
3	arc3	arc2	0,030834	0,009624	0,000270	0,008755	0,002785	0,007749	0,015132	2,605124	98,566055	1	100
4	arc3	arc4	0,108733	0,037768	0,000000	0,032148	0,014334	0,028205	0,057626	2,605124	92,934050	1	100
5	arc4	arc2	0,131505	0,038754	0,000327	0,034345	0,010753	0,025848	0,053031	2,803196	106,060217	1	100
6	arc4	arc3	0,127939	0,042007	0,005605	0,034045	0,011335	0,039591	0,061242	2,803196	107,603187	1	100
7	All	All	0,094233	0,029259	0,001034	0,025210	0,008686	0,022959	0,043325	2,683781	100,150729	1	100

Figure 4: Example of the outcome of the analysis of the differences of three versions of a data-set (arc1, arc2 and arc3). Q25, Q50 and Q75 are the 25% fractiles. Please note that the data in this figure are not those used for the tests later on in the paper.

2.4. Application of Monte Carlo simulation of uncertainty of area calculation

To simulate the effect of the (normal distributed) inaccuracy based on the inaccuracy model (see above) on individual vertexes a gaussian, random-number generator was needed. The option used for the present work was found using the GAUSS-function of the object-oriented language Python (ver. 2.3.4). The GAUSS-function returns a normal-distributed random-number, given a mean and a standard deviation. The application was made prior to the release of ArcGIS ver. 9.0, so the direct linkage between Python and geo-features was not available.

² The AML is available from the author.

Accordingly – as a very simple solution – the python-programme was working on generate-files³.

For each vertex of the ‘original’ polygon new x- and y-coordinates were calculated, using the original coordinate values as mean and the standard deviations obtained from the inaccuracy model. For each scale 50 copies were produced. Examples of 5 copies for the simulation in 1:10.000 are shown together with the ‘original’ in figure 5.

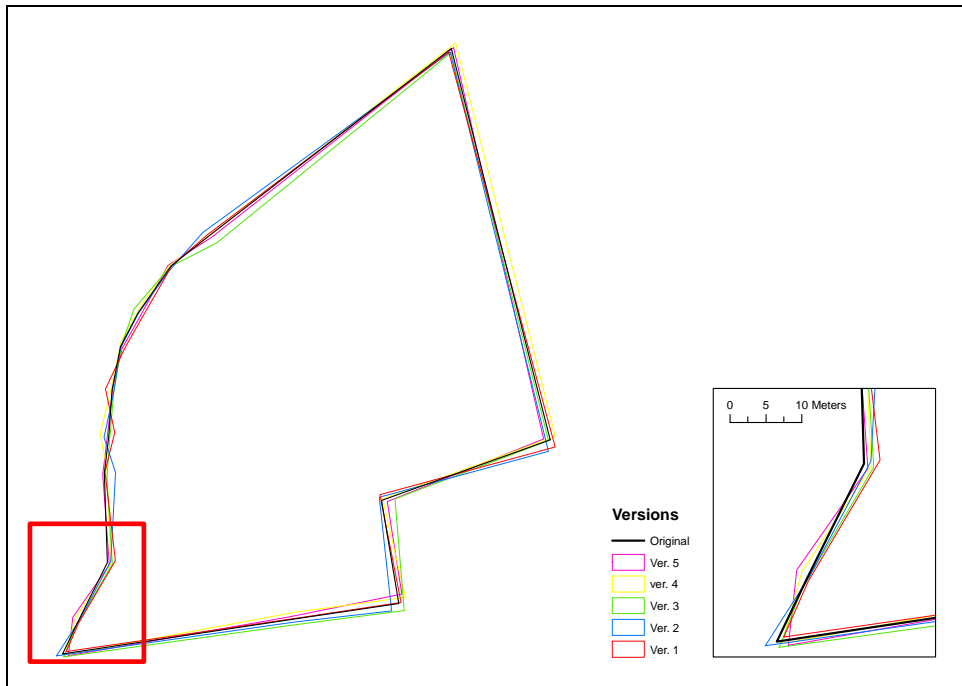


Figure 5: . Examples of 5 copies for the simulation in 1:10.000 is shown together with the ‘original’. Simulations are based on a standard deviation of 1.805564, based on the repeated copies (see above).

3. Results

3.1. Application of the ‘classical’ method of assessment of area-calculation

Results from the ‘Classical’ assessment method at different scales are show in table 4. The ‘original’ polygon was 29532,47 m².

Table 4: Results of the assessment of area calculations using the ‘classic’, buffer-method.

Scale	0.5 mm	Area of buffer	% of ‘Original’
1:2.000	1 m	1515.21	5.13
1:10.000	5 m	7538.53	25.53
1:25.000	12.5 m	18669.69	63.22

³ Readable/writable by the GENERATE/UNGENERATE commands of Workstation Arc/INFO.

3.2. Construction of an uncertainty model and application of Monte Carlo simulation of uncertainty of area calculation

Running the inaccuracy assessment (AML-) procedure provides statistics for spatial deviations of all coverages (or shape-files) vs. all. In the example displayed in figure 6, five coverages (blok0 through 4) are used. Blok0 represents the 'original' polygon, blok1 through 4 represents four different digitized versions. In the present case it is only the standard deviations of the digitized versions and the 'original' that are of interest. The average of the standard deviations of these four events (marked with a frame in figure 6) is used as the general estimate: $(2.02+1.85+1.41+1.93)/4= 1.81$ in the present case (Scale: 1:10.000).

IN	OUT	MIN	MAX	MEAN	STD	Q20	Q40	Q60	Q80
blok0	blok1	0,13	6,84	2,48	2,02	0,56	1,75	2,15	4,45
blok0	blok2	0,22	6,84	2,40	1,85	0,88	1,25	3,15	3,43
blok0	blok3	0,01	5,11	1,39	1,41	0,47	0,72	1,06	2,54
blok0	blok4	0,08	6,84	2,17	1,93	0,66	1,20	1,94	4,06
blok1	blok0	0,00	4,54	1,08	1,09	0,28	0,49	1,30	1,51
blok1	blok2	0	5,94	1,94	1,72	0,56	1,00	1,99	2,88
blok1	blok3	0	3,90	1,34	1,42	0,31	0,51	0,79	3,30
blok1	blok4	0	5,42	1,20	1,48	0	0,54	1,16	1,89
blok2	blok0	0,07	3,70	1,45	1,23	0,39	0,49	1,89	2,90
blok2	blok1	0	6,13	2,37	1,61	0,94	2,37	2,78	3,24
blok2	blok3	0,01	4,65	1,79	1,50	0,61	0,79	2,63	3
blok2	blok4	0	5,84	2,48	1,57	1,19	2,03	2,66	3
blok3	blok0	0,04	5,00	1,22	1,25	0,28	0,79	1,05	2,08
blok3	blok1	0	5,86	1,99	1,88	0,28	0,69	2,63	3,49
blok3	blok2	0,49	5,31	2,37	1,69	0,56	1,27	3,09	3,70
blok3	blok4	0	5,31	1,63	1,59	0	0,65	1,47	2,94
blok4	blok0	0,02	3,98	1,42	1,07	0,45	0,95	1,86	2,09
blok4	blok1	0	5,53	1,29	1,66	0	0,53	0,84	2,61
blok4	blok2	0	5,22	2,37	1,49	0,60	1,99	2,66	3,43
blok4	blok3	0	3,29	1,51	1,02	0,79	1,08	2,07	2,21
All	All	0,05	5,26	1,79	1,53	0,49	1,05	1,96	2,94

Figure 6 example of result from assessment of the digitizing uncertainty of the map in 1:10.000.

The standard deviations calculated this way for all three scales (1:2.000, 1:10.000 and 1:25.000. See tables 5) were used for the simulation of 50 versions of the original block. For each scale the average and standard deviation of the area was calculated.

Table 5: Result from the Monte Carlo Simulation of area calculations.

Scale	Mean of Vertexes differences	STD of Vertex differences	Mean Area	STD Area	99 % conf. intv ⁴ . (m ²)

⁴ One-sided, t-distributions, DF = 50.

1:2.000	0.46	0.30	29526.09	65.74	157.97
1:10.000	2.10	1.81	29495.92	293.83	706.07
1:25.000	3.84	2.80	29509.05	472.29	1134.91

Finally in table 6 the magnitude of the inaccuracy in area-calculation for the classical vs. the Monte Carlo method are compared. It appears that the classical method arrives at ranges of inaccuracy between 9.59 and 16.45, as large as the proposed Monte Carlo method.

Table 6: Ranges of inaccuracy in area calculation as found by the two methods used.

Scale	'Classic' method	Monte Carlo Method	Ratio
1:2.000	1515.21	157.97	9,59
1:10.000	7538.53	706.07	10,68
1:25.000	18669.69	1134.91	16,45

4. Conclusions and perspectives

One of the purposes of this paper has been to propose a means for simulation of inaccuracy in area calculation, propagated through the data-automation process (digitizing). This was obtained by Monte Carlo simulation using a gaussian model of the inaccuracy of data-capture of individual vertexes. Parameters for the inaccuracy model were obtained from repeated digitization in different scales.

Future work, using the methods and schemas proposed will include setting up a comprehensive test for different types (different degrees of complexity) of figures of different scales. Area accuracy of given figures can then be assessed, given their complexity and the digitizing scale.

Even for a single theme, on a single map sheet, conditions for (the quality of) data capture can vary. E.g. when digitizing a soil map, inaccuracy of demarcations between soil-types can vary from soil-type to soil-type: The clay-sand boundary may be pretty well defined, whereas boundaries between loamy and clayey soil-types might be less well defined. In that case inaccuracy models have to be defined and applied individually for different types of demarcations.

References

Zhang, J. and Goodchild, M. 2002. *Uncertainty in Geographical Information*. Taylor & Francis. Pp. 288.

Virrantaus, K. 2003. Post Graduate Course 'Uncertainty in Geographical Information' by Michael Goodchild. Hosted by Prof. Kirsi Virrantaus at Helsinki University of Technology. e-mail: kirsi.virrantaus@hut.fi. phone: +358 9 4513912