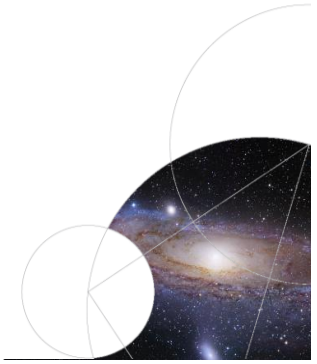
Quality and accuracy

Introduction to GIS



Hans Skov-Petersen (hsp@ign.ku.dk)

Guess what Chat GPT says
What are the major problems with present day GIS?

Quality and accuracy of geodata (I)? Can you be a little more specific?

- **Accuracy** is the deviation between reality and a measurement
  - In principle, it **cannot be assessed!**
  - In practice, accuracy is assessed by **comparing with 'better' or repeated measurements** of 'ground truth'
- **Precision** describes the capabilities of the instruments and methods used.
  - It can be assessed as the **deviation between repeated measures** of the same object by the same method
  - With respect to he captured data (measurements) precision applies to the **number of digits** presented by a given (computer-) registration.
- **Completeness** is a measure of the extent or amount of a give phenomena that was captured by an automation process.
- **Resolution** is the granularity – the smallest objects or characteristics you can possible reveal from a dataset.
- **Compatibility** describes the extent to which data from different sources can be used together due to diversity definitions and quality in
  - Space - e.g. different scales
  - Time
  - Semantics, characteristics (attributes)
- **Consistency** addresses potential differences in production method and quality – like compatibility - but across a single data set

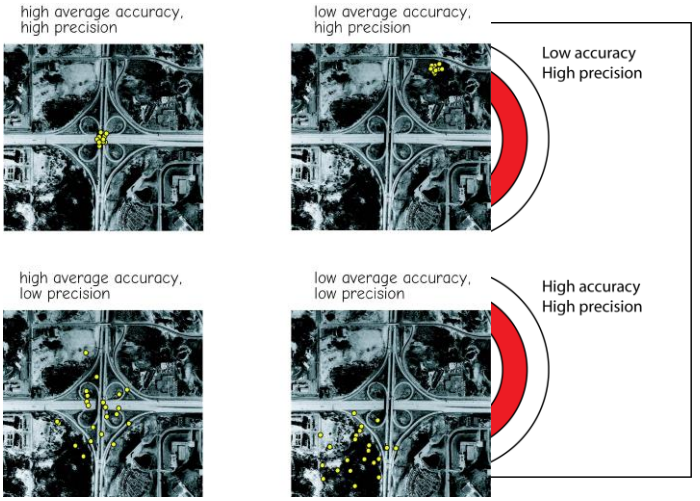
UNIVERSITY OF COPENHAGEN

Quality and accuracy of geodata (II)?

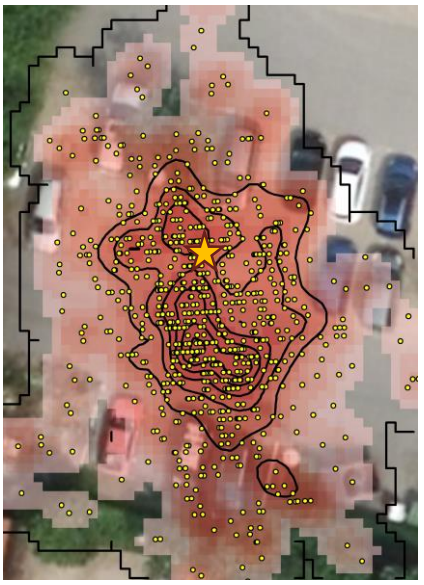
- **Bias** is the systematic variation of data from reality
  - Data production method
  - Operator's preferences, education, knowledge
  - Instruments lacking calibration
  - Shifts in projection, datum etc.
- **Applicability/appropriateness/suitability** for specific analysis or cartographic presentations
- **Fitness for use.** 'Good quality' is not an objective norm.
  - The quality, accuracy, precision ect. of a data set can only be judged in relation to its use
  - … and user
- **Trustworthiness**. The collective of all above: To which degree can you believe what you see – or can reveal from what you see - and make the right decision.

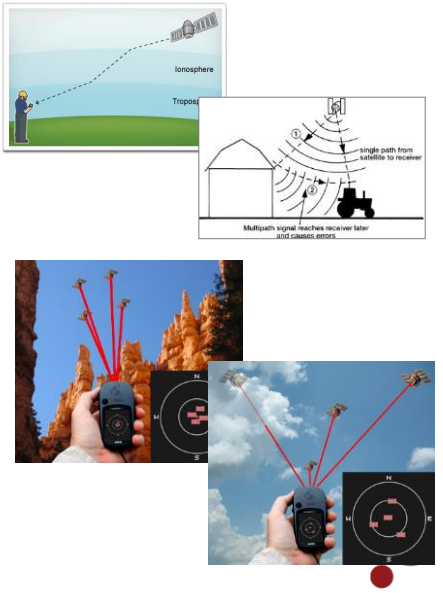
Accuracy and Precision

- Accuracy is how well data reflects reality
- Precision reflects the distribution of repeated measurements



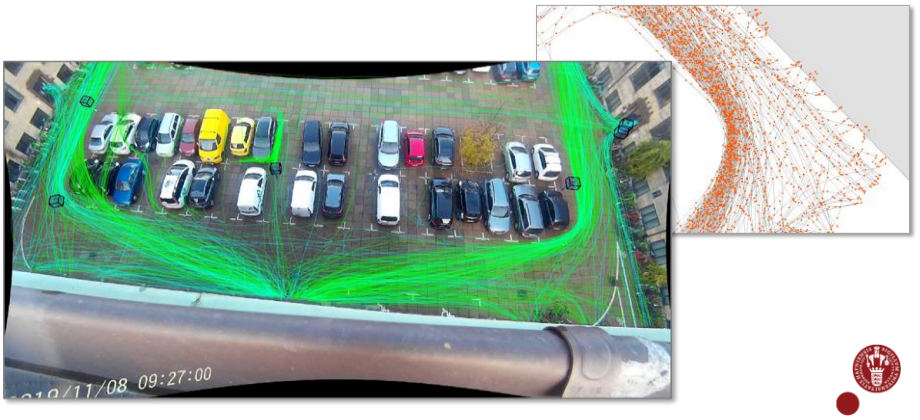
Assessment of precision
… of smart phone GPS location





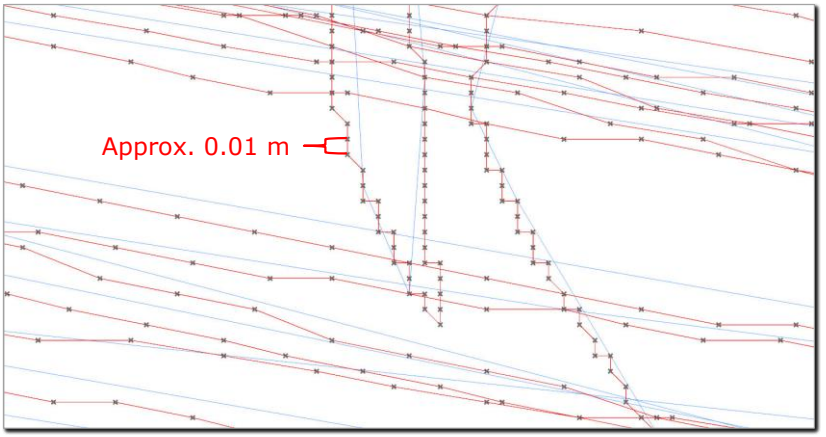
Assessment of precision
A note on the number of decimals
… let's take a look at camera-base registration

- Very high spatial and temporal resolution and accuracy (compared to GPS)

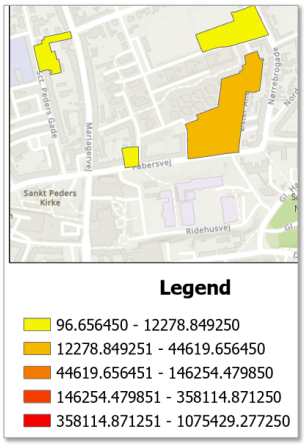


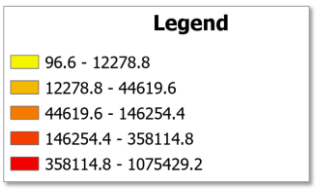
Camera-base registration

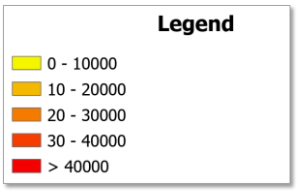
- Details at the edge of the reasonable…
- Why the edgy shapes?



Signalling precision
Another note on the number of decimals
… how do you label your legends?







Accuracy assessment (points)

According to the US 'National Standard for spatial Data Accuracy (NSSDA), Bolstad p 617:

- Identify/sample test points
- Identify their 'true' location (by higher order accuracy – measurements or existing, alternative data)
- Calculate mutual spatial deviations (distance)
- Calculate error statistics
  - Mean, STD and 95% confidence (assuming normal distribution)

Standards (rules of thumbs) of accuracy

'National Standard for spatial Data Accuracy' (NSSDA), Bolstad p 617:

ID	x (true)	x (data)	x differ-ence	(x differ-ence)^2	y (true)	y (data)	y differ-ence	(y differ-ence)^2	sum x diff^2 + y diff^2
1	12	10	2	4	288	292	-4	16	20
2	18	22	-4	16	234	228	6	36	52
3	7	12	-5	25	265	266	-1	1	26
4	34	34	0	0	243	240	3	9	9
5	15	19	-4	16	291	287	4	16	32
6	33	24	9	81	211	215	-4	16	97
7	28	29	-1	1	267	271	-4	16	17
8	7	12	-5	25	273	268	5	25	50
9	45	44	1	1	245	244	1	1	2
10	110	99	11	121	221	225	-4	16	137
11	54	65	-11	121	212	208	4	16	137
12	87	93	-6	36	284	278	6	36	72
13	23	22	1	1	261	259	2	4	5
14	19	24	-5	25	230	235	-5	25	50
15	76	80	-4	16	255	260	-5	25	41
16	97	108	-11	121	201	204	-3	9	130
17	38	43	-5	25	290	288	2	4	29
18	65	72	-7	49	277	282	-5	25	74
19	85	78	7	49	205	201	4	16	65
20	39	44	-5	25	282	278	4	16	41
21	94	90	4	16	246	251	-5	25	41
22	64	56	8	64	233	227	6	36	100
								Sum	1227
								Average	55.8
								RMSE	7.5
								NSSDA	12.9

Table 14.4

Sqrt(55.8)

Multiplying by 1.73

SQR_Dist	Dist
20	4.5
52	7.2
26	5.1
9	3.0
32	5.7
97	9.8
17	4.1
50	7.1
2	1.4
137	11.7
137	11.7
72	8.5
5	2.2
50	7.1
41	6.4
130	11.4
29	5.4
74	8.6
65	8.1
41	6.4
41	6.4
100	10.0

Order		SQR_Dist	Dist
1	5%	2	1.4
2	9%	5	2.2
3	14%	9	3.0
4	18%	17	4.1
5	23%	20	4.5
6	27%	26	5.1
7	32%	29	5.4
8	36%	32	5.7
9	41%	41	6.4
10	45%	41	6.4
11	50%	41	6.4
12	55%	50	7.1
13	59%	50	7.1
14	64%	52	7.2
15	68%	65	8.1
16	73%	72	8.5
17	77%	74	8.6
18	82%	97	9.8
19	86%	100	10.0
20	91%	130	11.4
21	95%	137	11.7
22	100%	137	11.7

Average	55.8
RMSE	7.5
95%	12.9

The distance within which 95% of the tested points are located (assuming normal distribution)

Rules of thumb

The US National Map Accuracy Standard (NMAS) (Bolstad, p 616) states that:

- 90% of points
- must be within 1/30 of an inch (2.54 cm) = 0.9 mm
- for maps in scale 1:20.000 or larger (means more detailed)

A similar Danish rule (not formally documented) says that (or at least, that's what I normally go with):

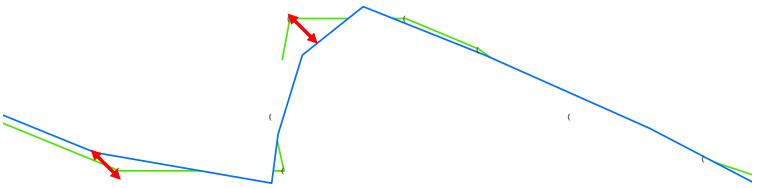
- 90% of the points
- Must be within plus/minus 0.5 mm
- For all map types

	± 0,25 mm	± 0,5 mm	± 0,75 mm	± 1 mm
Scale	m	m	m	m
1:25.000	6,25	12,50	18,75	25,00
1:50.000	12,50	25,00	37,50	50,00
1:100.000	25,00	50,00	75,00	100,00
1:200.000	50,00	100,00	150,00	200,00
1:500.000	125,00	250,00	375,00	500,00
1:1.000.000	250,00	500,00	750,00	1000,00
1:2.000.000	500,00	1000,00	1500,00	2000,00

UNIVERSITY OF COPENHAGEN

Assessment of accuracy (lines)

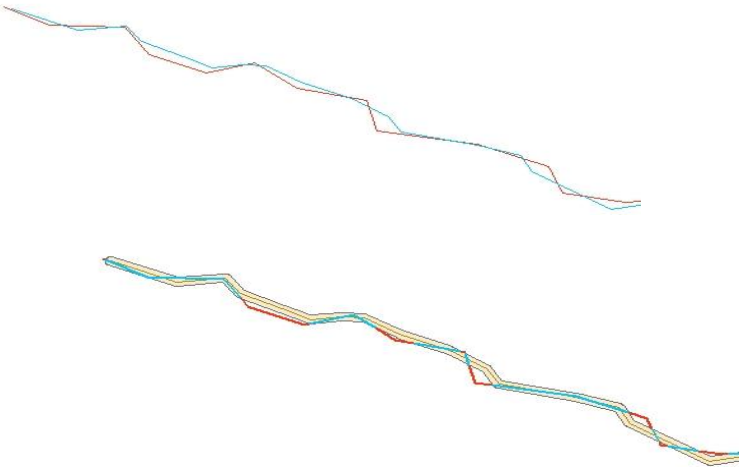
- In case of digitizing **linear data** - where vertexes cannot be mutually compared - the orthogonal distance from vertexes on one version of a line to another can be used



UNIVERSITY OF COPENHAGEN

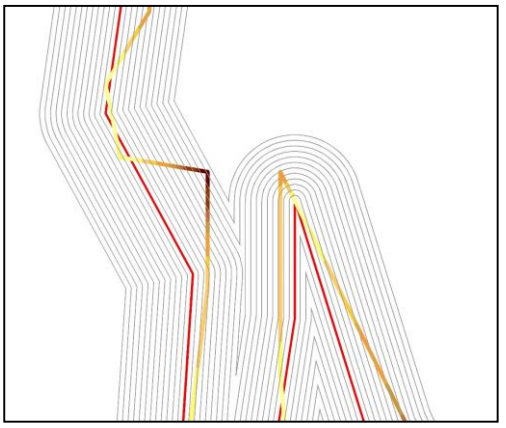
Assessment of accuracy (lines)

Or, you can assess the overlay of a buffer – of one line – vs another line …



Assessment of accuracy (lines)

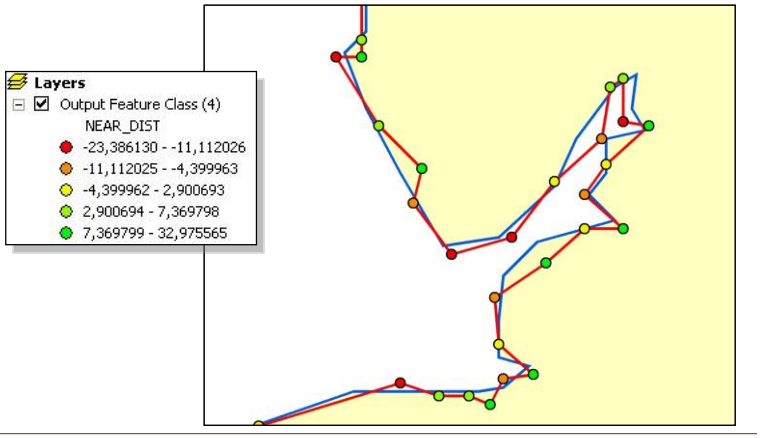
- Multi Ring Buffers (which will be the approach in the exercise)



UNIVERSITY OF COPENHAGEN

Assessment of accuracy (lines)

- A problem is that the deviation is two-sided – positive to one side, negative to the other
- If distances are assessed this way an assumption of normal distribution would be easier to justify



Accuracy assessment (polygons)

Let's say that you were to map the distribution of Rosa Rogusa of a dune area…

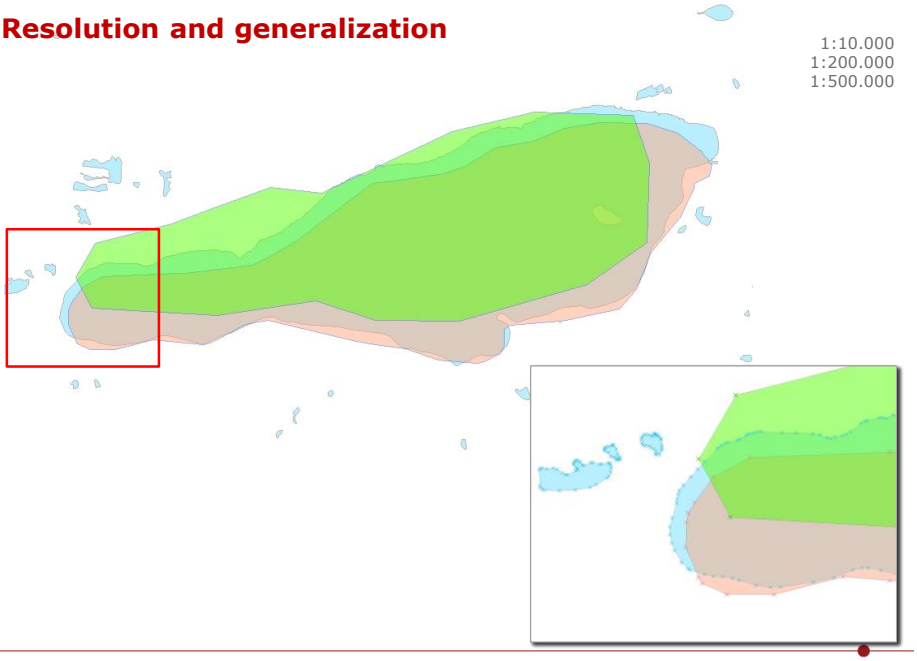
How would you approach and document the data you produce?

UNIVERSITY OF COPENHAGEN IGN

Pause

Resolution and generalization

1:10.000
1:200.000
1:500.000



Resolution and the length of the Danish coastline

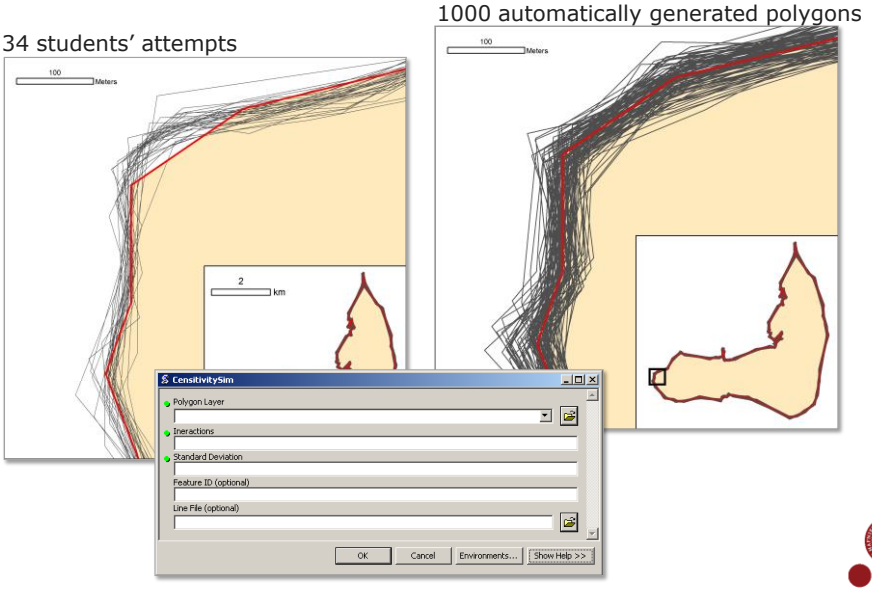
On the map: 6.637 Km
- … in 1: 500.000
- 1:200.000: 7.113 Km
- 1:10.000: 8.472 Km

Officially (2014)
- 1:100.00: 7.314 Km
- 1:10.000: 8,750 Km

Assessment of accuracy of derived values (length and area)

34 students' attempts

1000 automatically generated polygons

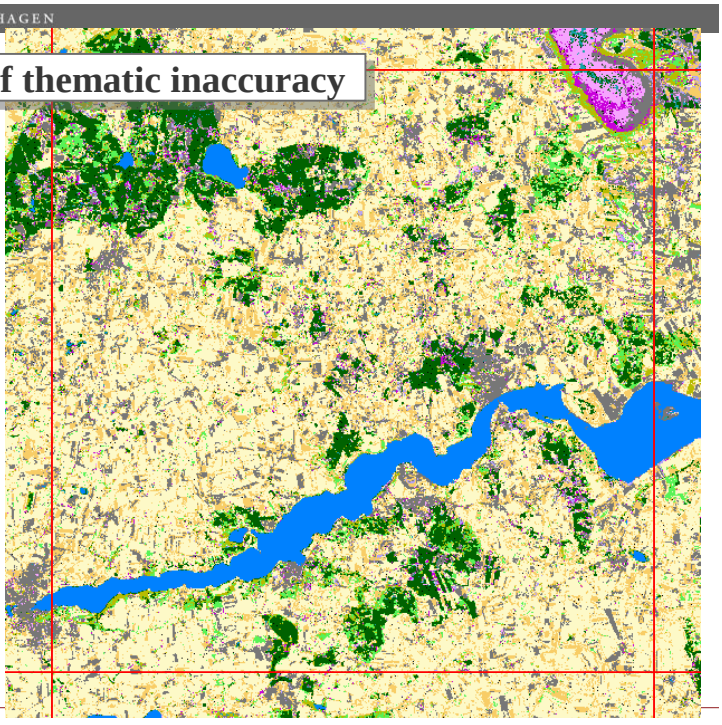


Assessment of thematic inaccuracy

The case of the Area Information System (AIS)

Land Use

Land Cover



Thematic accuracy: Confusion matrix (1)

AIS Land cover — Grass land — Coniferous forest

AIS Land use — Agriculture

	X0	X1	X3	X5	X7	X8	X10	X11	X14	X15	X16	X18	Sum
0	44												220
1122	1		7	3	2	5	5	2	1	1	1	3	1.646
1123	1		4	3	3			2	1	1	1	11	2.916
1128												1	279
1223			1	1	1	1	1	1				1	441
1224			2	2	2	1	1	1			1	2	865
1226													56
2112	36		60	66	74	15	43	42	51	40	30	45	32.118
2300													100
2310													28
3110	3		1	1	4	4	2	4	9	19	14	7	3.931
3120	10		8	13	7	45	12	26	22	30	40	12	9.260
3130													22
3210			1	1	1	1		1	1	1	1	3	887
3220	1		1	5		1	2	5	2	2	3	2	1.012
3310													7
3330												1	219
4110	1			1	2	2			1	1	3	3	983
4112				2	1			2				2	655
4120	2	96	2	2	2	9	30	12	5	4	4	4	4.685
4210		1			1	3	1						245
5120		3			1	9	1		4		1	1	622
5121													2
5126													21
5230			12			2							1.165
6000													116
Sum	100	100	100	100	100	100	100	100	100	100	100	100	0

AIS codes:
- Land use
- Land Cover

Thematic inaccuracy: Confusion matrix (2)

AIS Land cover

Barren land

AS Land Use — Agriculture — Forest — Sand dunes

	X0	X1	X3	X5	X7	X8	X10	X11	X14	X15	X16	X18	Sum
0	100												100
1122			40	2	15	2	3	2	4	2	3	26	100
1123			13	1	17			1	2	1	1	63	100
1128			9	1	11			1	6	2	3	68	100
1223	1		19	2	26	1	1	2	6	3	5	35	100
1224			17	3	28	1	1	1	3	2	3	42	100
1226			18	1	16	2	3	5	13	3	7	31	100
2112	1		17	3	35		1	2	8	6	5	23	100
2300			9	1	40				5	2	2	45	100
2310			15	2	41			2	4	3	2	30	100
3110			2		14	1	1	1	13	22	18	28	100
3120	1		8	2	12	3	1	4	12	15	21	21	100
3130	1		6		6	2	2	3	25	23	21	11	100
3210			5	2	22	1	1	2	5	5	4	54	100
3220			6	6	7	1	2	8	12	7	14	37	100
3310		2	82	1	1	3	2				1	7	100
3330			12	1	3			3		4		76	100
4110	1		1	1	27	1			3	3	14	48	100
4112			2	4	34		1	5	3	3	1	46	100
4120		49	5		5	1	6	4	5	4	5	15	100
4210		5	9	2	33	7	3	2	5	1	5	28	100
5120		11	4		13	9	1	1	37	2	8	14	100
5121			20	3	22		2	2	3	9		40	100
5126			12	1	25	4	4	1	7	2	24	19	100
5230			95			1					1	1	100
6000			14	1	13		1	5	7	5	6	48	100
Sum	505	2.405	9.277	1.262	15.236	630	970	1.460	5.205	4.473	4.886	16.192	116

AIS codes:
- Land use
- Land Cover

Completeness

Like in the case of accuracy/precision, completeness can only be assessed by either

- comparison with something regarded as being closer to the 'truth' or
- by repetition of *randomly selected portions* of data collection.

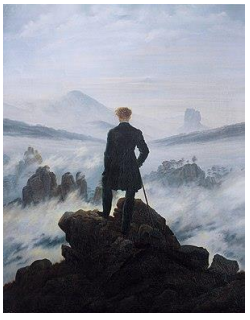
Completeness can be assessed in terms of e.g.

- *Counts* – the number of features within the test areas
- *Measures* – aggregated derived measures, e.g length and areas
- *Attributes* – amount of blank areas, confusion matrixes etc.

And so what?

We are at point in time where…

- Data production is hidden to its end-users
- Analytical techniques are well developed (and not a prime concern of most GIS-users)
- Storage and distribution of (massive amounts) of data is becoming increasingly important
- Dynamic – and often real-time – data are becoming more prominent
- Data and spatial analysis are used more and more in the (mass-) media
- … and we don't know who we are talking to



How about all our non-professional users? We need to talk….

In the ol' days…

any GIS project would be constituted by 85% of data collection.

We don't need that anymore, which is so great.

But… realizing that *'GIS is a mass-media'*, things become much more challenging.

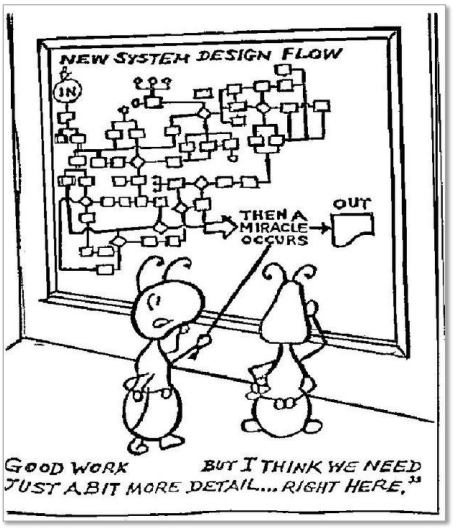
As professionals…

we have a fair idea of quality and what it means. We know to which extent it is 'correct' and how we can have faith in it.

It cannot be expected for **the average, unknown target group** of mass-media communication.

We, as professionals, have an obligation to enhance communication of quality and trustworthiness of the geodata presented to the public

© Niels Villum Petersen



There are facts … and alternative facts?

Talk to your neighbor.

How would you assess the **accuracy** and **precision** of measurements of the number of attendees at the two occasions?

How Kellyanne Conway ushered in the era of 'alternative facts'

Don't believe in any data you get...

They are nothing but representations
... of something
... for someone
... produced by someone
... for some purpose

Thanks
That's it for today

Hans Skov-Petersen
hsp@ign.ku.dk

UNIVERSITY OF COPENHAGEN IGN

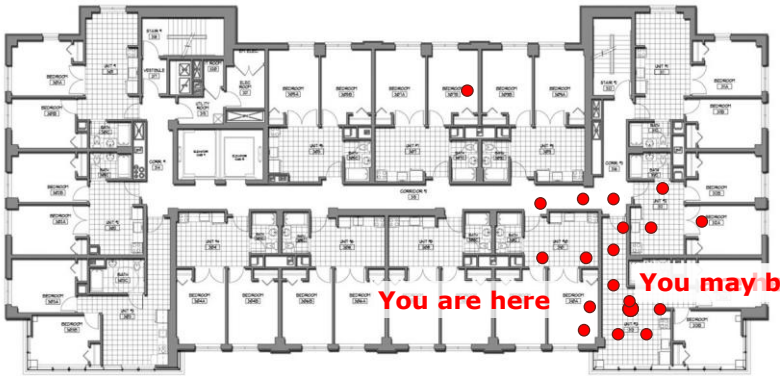
AIS Land Use

Kode	Beskrivelse
1100	Befæstet overflade
1110	Bykerne
1120	Lav bebyggelse
1121	Høj bebyggelse
1122	Åben bebyggelse
1123	Bebyggelse i åbent land
1210	Industri
1221	Motorvej
1222	Motortrafikvej
1223	Vej > 6 m
1224	Vej 3-6 m
1226	Jernbane
1228	Bro
1229	Dæmning
1240	Lufthavn
1242	Landingsbane
1310	Råstofområde
1340	Teknisk areal
1341	Kirkegård
1420	Sportsanlæg
1421	Rekreativt område
1422	Klippet græs
2112	Landbrug
2222	Gartneri
2300	Græsarealer
2310	Græs i byområder
2430	Blandet landbrug/natur
3100	Skov
3110	Løvskov
3120	Nåleskov
3130	Blandet skov
3210	Overdrev
3220	Hede
3250	Blandet natur
3310	Sand/klit
3330	Anden overflade med ringe vegetation
4110	Eng
4112	Vådområde
4120	Mose
4130	Strandeng
5120	Sø
5121	Vandløb > 8-12 m
5123	Sø-rørskov
5126	Dambrug
5230	Hav
6000	Uklassificeret

AIS land Cover

Tabel 3. Oversigt over Land Cover Map og Land Cover Plus

Celle-værdi	Klasser	Kortklasse LCM	Kortklasse LCM til LCP	Kort beskrivelse af klassen
0	Ukendt klasse	•		Ikke-klassificerede billeddata
1	Åbent vand	•		Områder med permanent åbent vand
3	Ubevokset overflade	•		Inkluderer naturligt ubevoksede overflader (f.eks. strande), ubevokset landbrugsjord og bebyggede områder
4	Vegetation påvirket af tidevand		1,3,7	Saltmarsk, alger og anden kystvegetation med tegn på oversvømmelse
5	Græsbevokset hedeområde	•		Tørre, semi-naturlige områder domineret af græsser, uden fjernelse af dødt, vegetativt materiale i større omfang
6	Permanent kort græs		7	Områder med permanent kort græs pga. hyppig græsning eller slåning
7	Afgræsset eller slået græs	•		Græsområder med lille akkumulering af dødt, vegetativt materiale – dog kan biomassen være enten høj eller lav
8	Engområde	•		Årligt våde, semi-naturlige områder domineret af græsser, med eller uden fjernelse af dødt, vegetativt materiale i større omfang
10	Busk- og græsbevokset hedeområde	•		Tørre, semi-naturlige områder domineret af en detaljeret mosaik af græsser og træagtige buske f.eks. lyng, revling, blåbær, enebær.
11	Buskbevokset hedeområde	•		Hovedsagelig tørre, semi-naturlige områder domineret af træagtige buske f.eks. lyng, revling, blåbær, enebær
14	Busk- og skovområde	•		Tørre eller våde områder domineret af buskbevoksning
15	Løvskov	•		Skove domineret af løvfældende træer
16	Nåleskov	•		Skove domineret af nåletræer
17	Urtebevokset moseområde		8	Sæsonbestemte våde, semi-naturlige områder med både græsser og anden urtebevoksning
18	Sæsonbestemt arealdække	•		Områder, der skifter markant mellem et vegetations-dækket og ikke-vegetationsdækket stadium, inklusive dyrkede arealer



The you are here map

… of the Heisenberg institute