L7: Uncertainty in geographic information

Introduction to final projects

Longley et al., 2005, **Geographic Information Systems and Science**:
- ch. 6: Uncertainty

Uncertainty in geographic information:

- What is uncertainty?
- U1: Uncertainty in the conception of geographic phenomena
- U2: Uncertainty in the measurement and representation of geographic phenomena
- U3: Uncertainty in the analysis of geographic phenomena
- Dealing with uncertainty

Introduction to the final project

What is uncertainty?

It is impossible to make a perfect representation of the world, so **uncertainty** about it is **inevitable**.

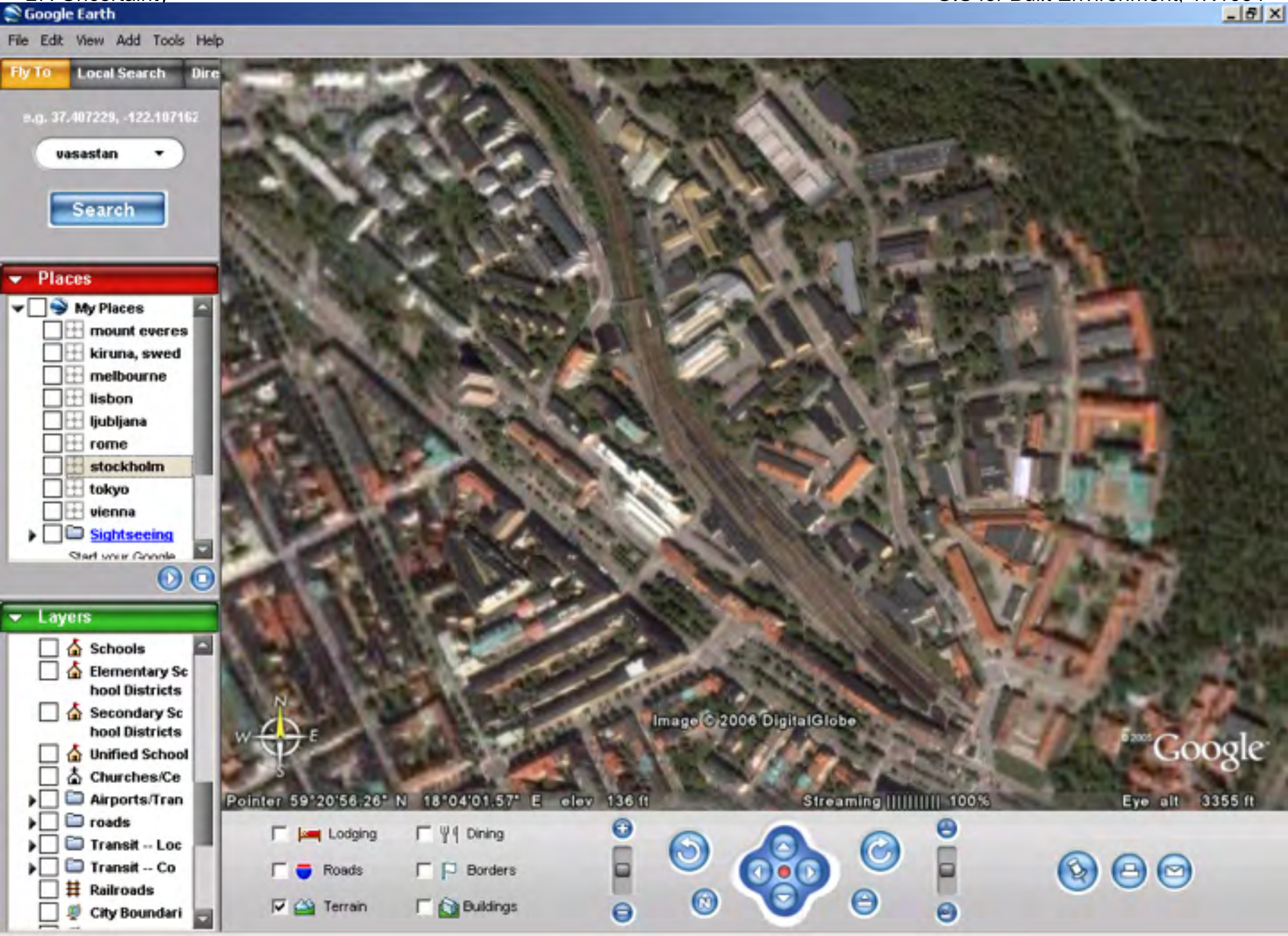
Geospatial data → wrongly measured

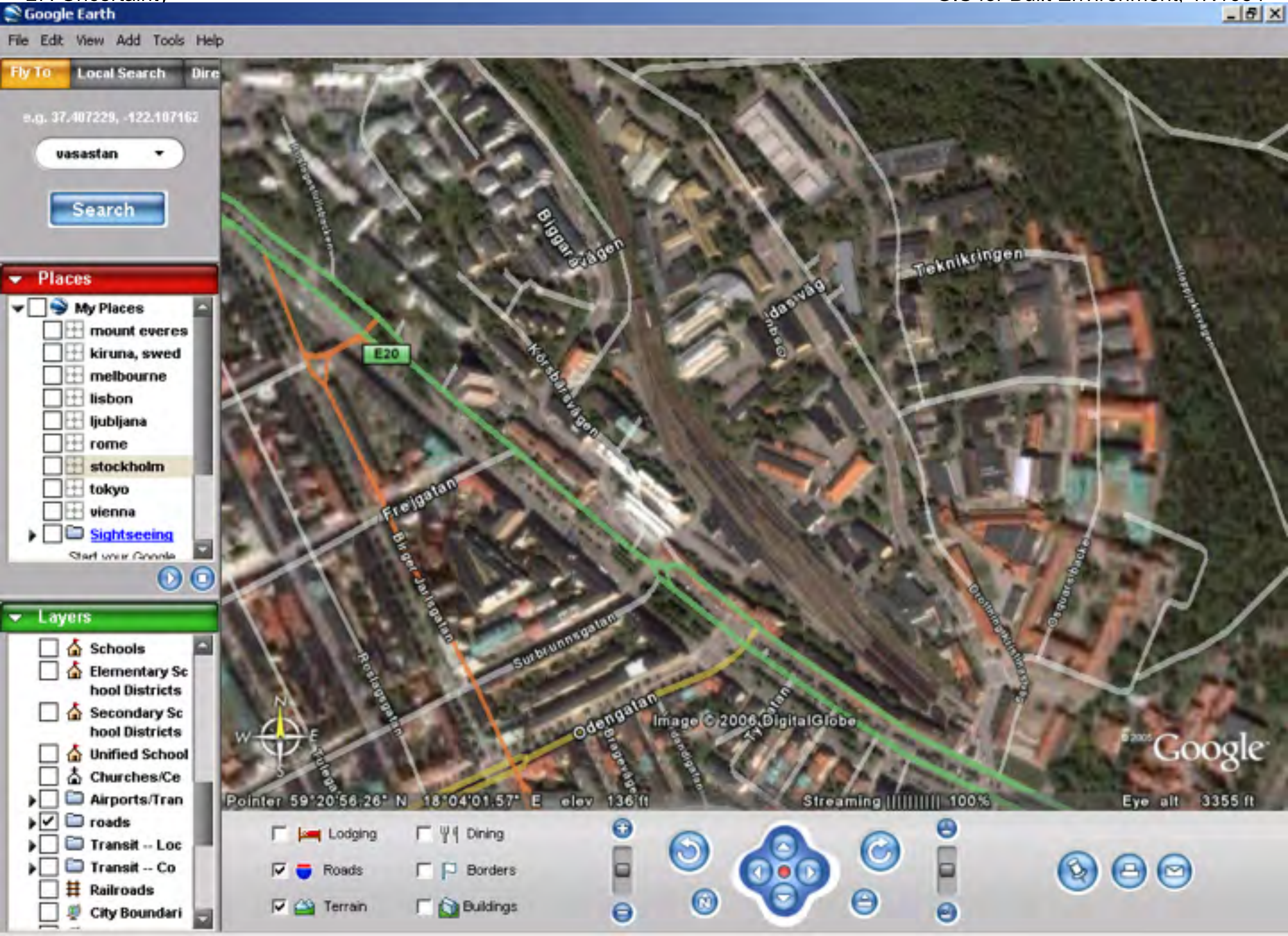
Geospatial data → out of date

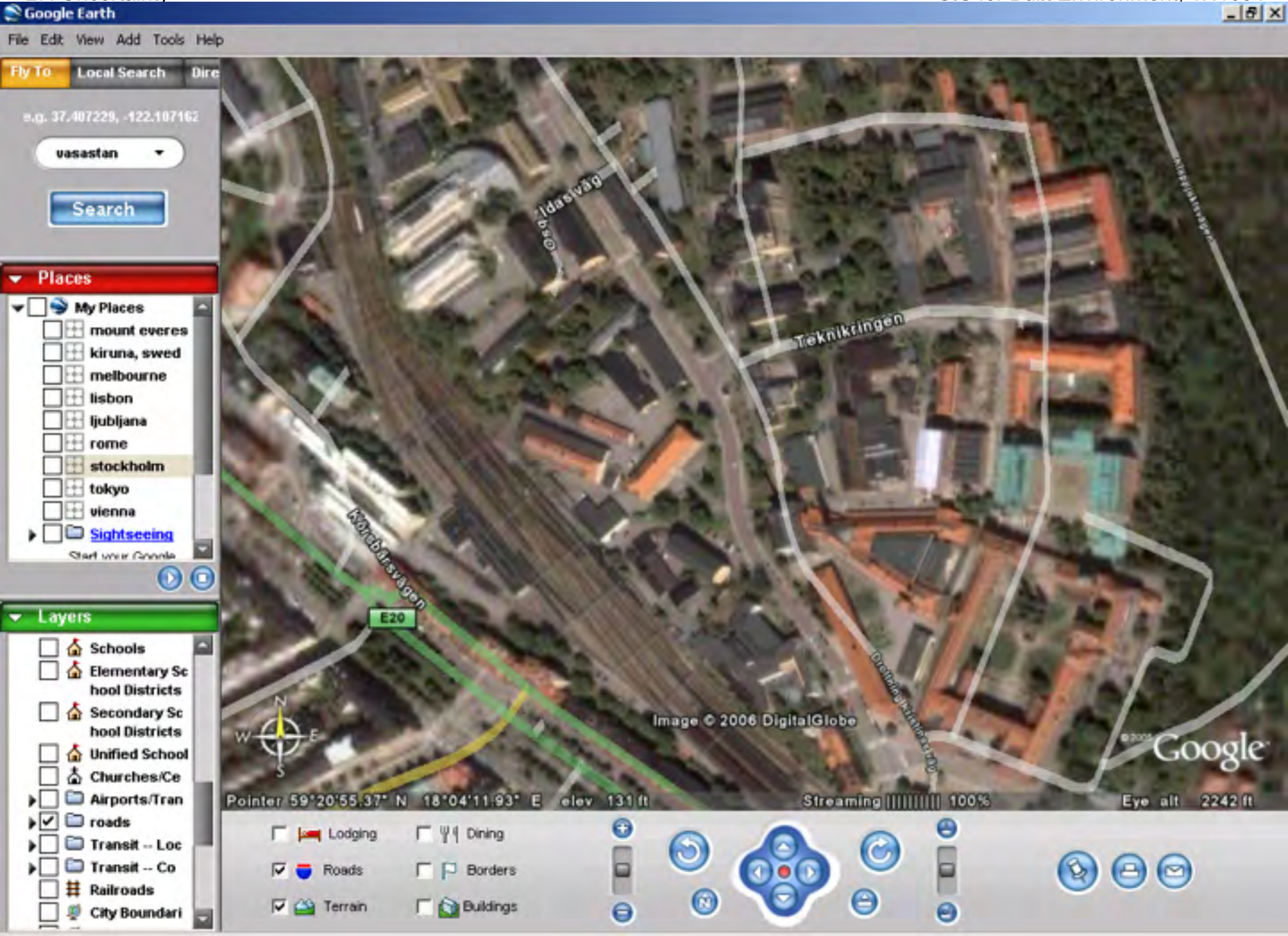
Geospatial data → too generalised

Geospatial data → wrong

Geospatial data → etc.


Google Earth
File Edit View Add Tools Help
Fly To
Local Search
Dire
e.g. 37.407229, -122.107162
vasastan
Search
Places
My Places
mount everes
kiruna, swed
melbourne
lisbon
ljubljana
rome
stockholm
tokyo
vienna
Sightseeing
Layers
Schools
Elementary School Districts
Secondary School Districts
Unified School
Churches/Ce
Airports/Tran
roads
Transit -- Loc
Transit -- Co
Railroads
City Boundari
Image © 2006 DigitalGlobe
Google
Pointer 59°20'56.26" N 18°04'01.57" E elev 136 ft
Streaming |||||||||| 100%
Eye alt 3355 ft
Lodging
Dining
Roads
Borders
Terrain
Buildings


Google Earth
File Edit View Add Tools Help
Fly To
Local Search
Search
Places
My Places
mount everes
kiruna, swed
melbourne
lisbon
ljubljana
rome
stockholm
tokyo
vienna
Sightseeing
Layers
Schools
Elementary School Districts
Secondary School Districts
Unified School
Churches/Ce
Airports/Tran
roads
Transit -- Loc
Transit -- Co
Railroads
City Boundari
Biggarvägen
Teknikringen
E20
Körsbärsvägen
Freigatan
Birger Jarlsgatan
Surbrunnsgatan
Odengatan
Roslagsgatan
Image © 2006 DigitalGlobe
Google
Pointer 59°20'56.26" N 18°04'01.57" E elev 136 ft
Streaming 100%
Eye alt 3355 ft
Lodging
Dining
Roads
Borders
Terrain
Buildings


Google Earth
File Edit View Add Tools Help
Fly To
Local Search
Dire
e.g. 37.407229, -122.107162
vasastan
Search
Places
My Places
mount everes
kiruna, swed
melbourne
lisbon
ljubljana
rome
stockholm
tokyo
vienna
Sightseeing
Layers
Schools
Elementary School Districts
Secondary School Districts
Unified School
Churches/Ce
Airports/Tran
roads
Transit -- Loc
Transit -- Co
Railroads
City Boundari
Teknikringen
Körbadsvägen
E20
Image © 2006 DigitalGlobe
Google
Pointer 59°20'55.37" N 18°04'11.93" E elev 131 ft
Streaming |||||||||| 100%
Eye alt 2242 ft
Lodging
Dining
Roads
Borders
Terrain
Buildings

Data quality

Every product of data integration has a certain level of **uncertainty** due to:

- Mistakes in original data
- Data processing

Data product is usable only when a certain **level of reliability** is reached.

Important: to present information about the quality of original data and the uncertainty from the processing steps to the user.

Aspects of data quality:

- **Lineage** – when the data was collected, what processing was used, etc.

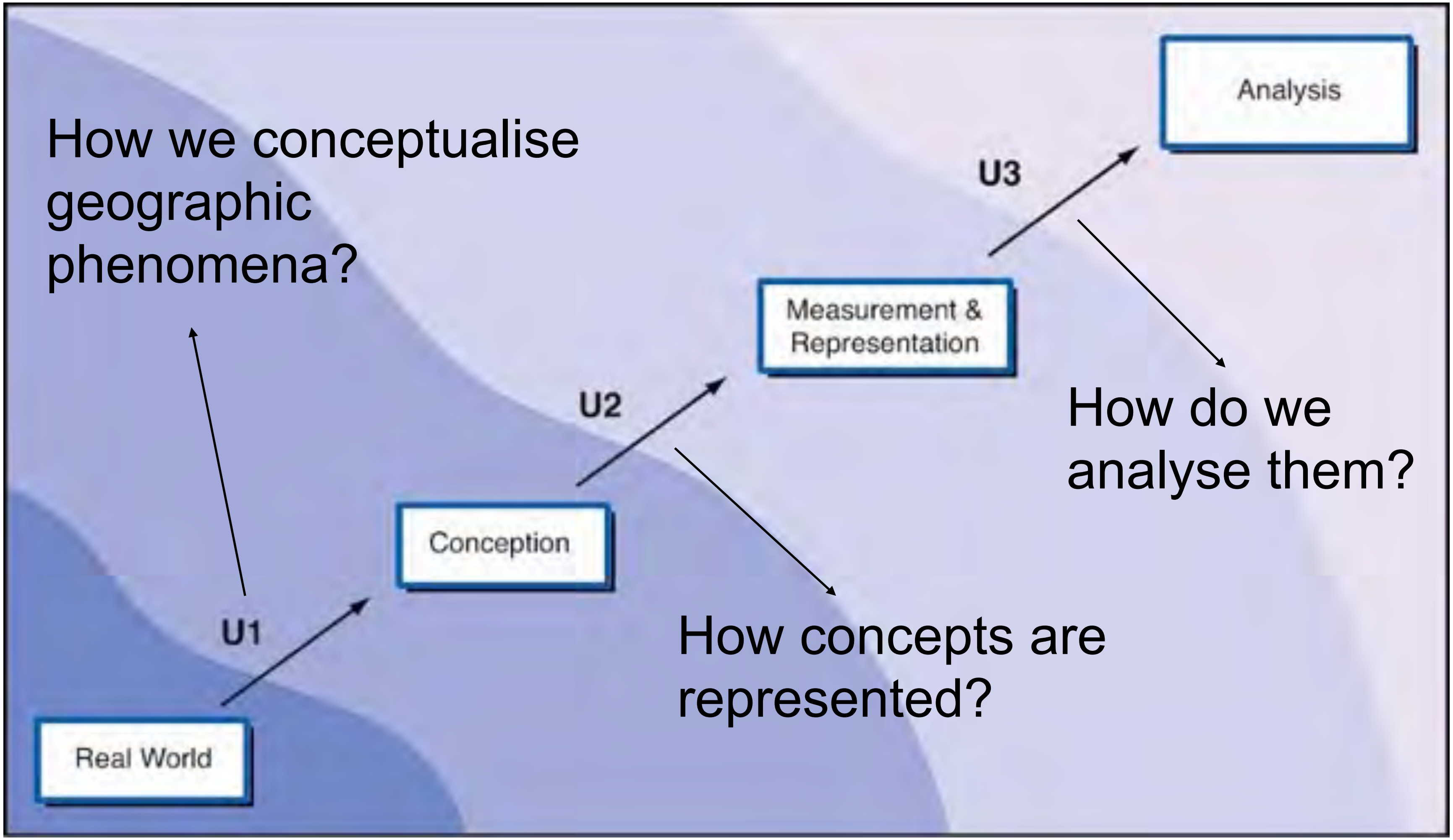
- **Positional accuracy** – how far is an object from its real position

- **Attribute accuracy** – what is the accuracy of attributes' values for an object

- **Logical consistency** – do the lines intersect in a point, are the areas closed polygons, etc.

- **Completeness** – is the data complete for the whole collection area

Where does uncertainty occur?



U1: Uncertainty in concepts

Uncertainty in concepts

Natural geographic units

Vagueness

Ambiguity

Natural geographic units

What is the natural unit for:
- soil profile
- land use
- a pocket of high unemployment
- a cluster or cancer cases
- environmental spillage from an oil tanker
- etc.

At what scale to investigate

relationship between radiation and lung cancer?

relationship between labour-force qualifications and unemployment?

Different units

Different scales

Different results

Vagueness

related to object's

boundaries

Where does one soil type end and another one start?

Is the defining boundary of a zone crisp and well-defined?

attributes

Is our assignment of a particular label to a given zone robust and defensible?

What number/% of oak trees is enough to define a landuse category *oak forest*?

implications

cognitive

Do people understand labels correctly?

cartographic

How to graphically portray vague boundaries?

statistical

Confidence for each boundaries/labels?

Ambiguity

Using indirect indicators

Direct indicators

a clear correspondence with a mapped phenomenon

household income

Indirect indicators

the best available measure is a perceived surrogate link with the phenomenon of interest

Rate of multiple cars ownership = Indirect indicator of household income

Cultural/language issues

Many objects are assigned different labels by different national or cultural groups, and such groups perceive space differently.

Object names and the topological relations between them are ambigious.

In Swedish:

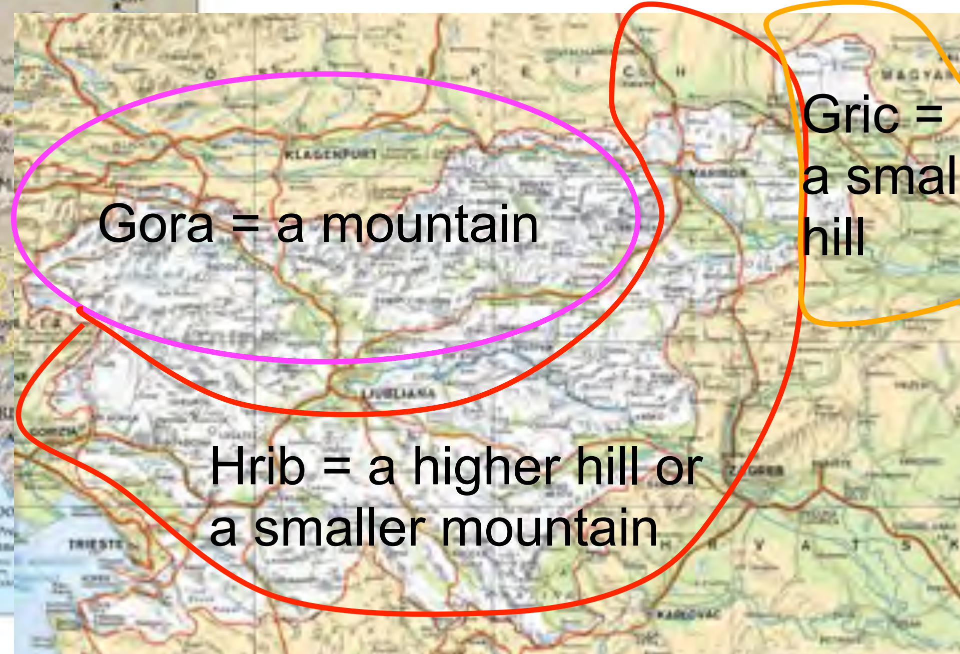
ett fjäll = a mountain in northern Sweden

ett berg = any other mountain

Language ambiguity

How do you define a mountain?

In Slovenian:



Gora = a mountain

Hrib = a higher hill or a smaller mountain

Gric = a smaller hill

Fuzzy approaches

Solving assignment process

↓

Probabilistic Interpretation by using **the fuzzy set theory**

In fuzzy set theory, it is possible to have partial membership in a set:

- membership can vary, e.g. from 0 to 1
- this adds a third option to classification: yes, no, and maybe (and to which extent).

↓

Dealing with sets that are not precisely defined, and for which it is impossible to establish membership cleanly.

Soil maps

Fuzzy model for imprecise soil polygon boundaries

Soil mapping:
- soil type classification
- defining soil polygon boundaries

→ sparse sampling in Finland due to large extent of the country.

↓

Defining soil boundaries in Finland between actual sampling sites:
manual interpretation by soil suveyors using
areial photos
geological maps
topographical map
knowledge about geomorphology

↓

Different surveyors produce different maps.

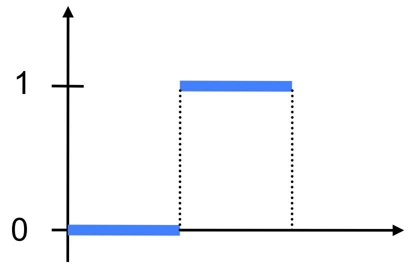
R. Sunila, E. Laine and O. Křemenova, Fuzzy model and kriging for imprecise soil polygon boundaries, Proc. of 12th International conference Geoinformatics 2004, Gävle, Sweden, June 2004.

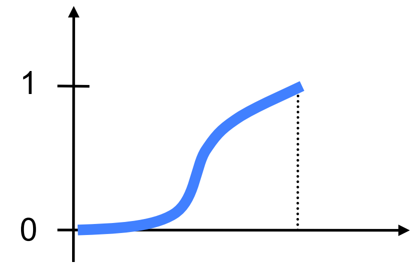
Fuzzy modelling for imprecise soil boundaries

Soil boundaries are not crisp in real world.

Fuzzy modelling is a way to take into account the gradual change between soil types and represent the uncertainy in soil polygon mapping.

Crisp 0/1 membership function

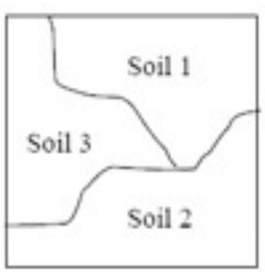


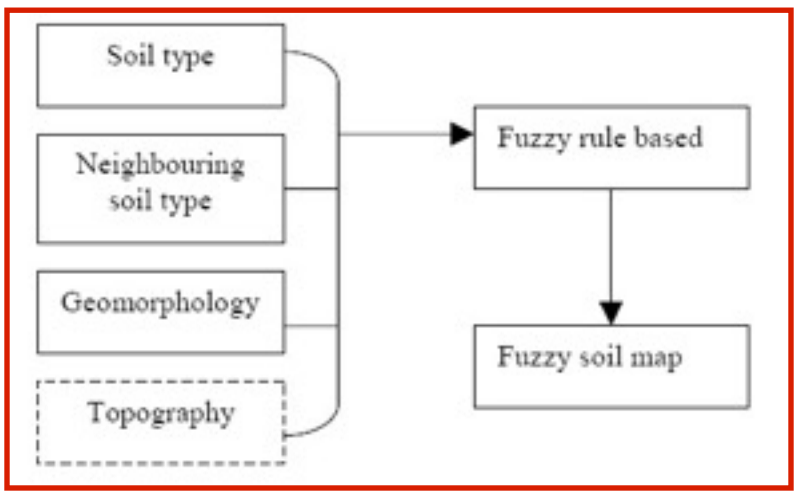


Fuzzy membership function

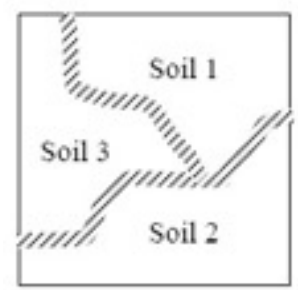
Fuzzy expert system

Soil map





Fuzzy map of the imprecision zone in the soil boundaries



Case study – Vampula area

Definitions of membership functions for fuzzy modelling based on specific soil characteristics of the study area.
Rules for:
- transitional zone -> varies according to soil type
- zone width -> based on expert knowledge

Original soil map

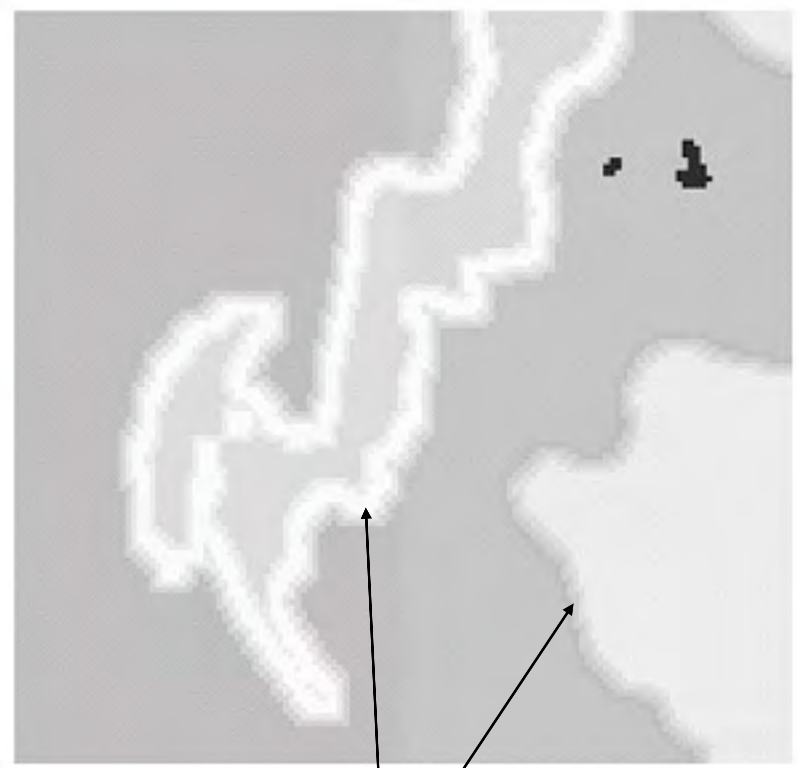
25m resolution grid, 3.2 km^2

Approximate location of the case study area

Original map

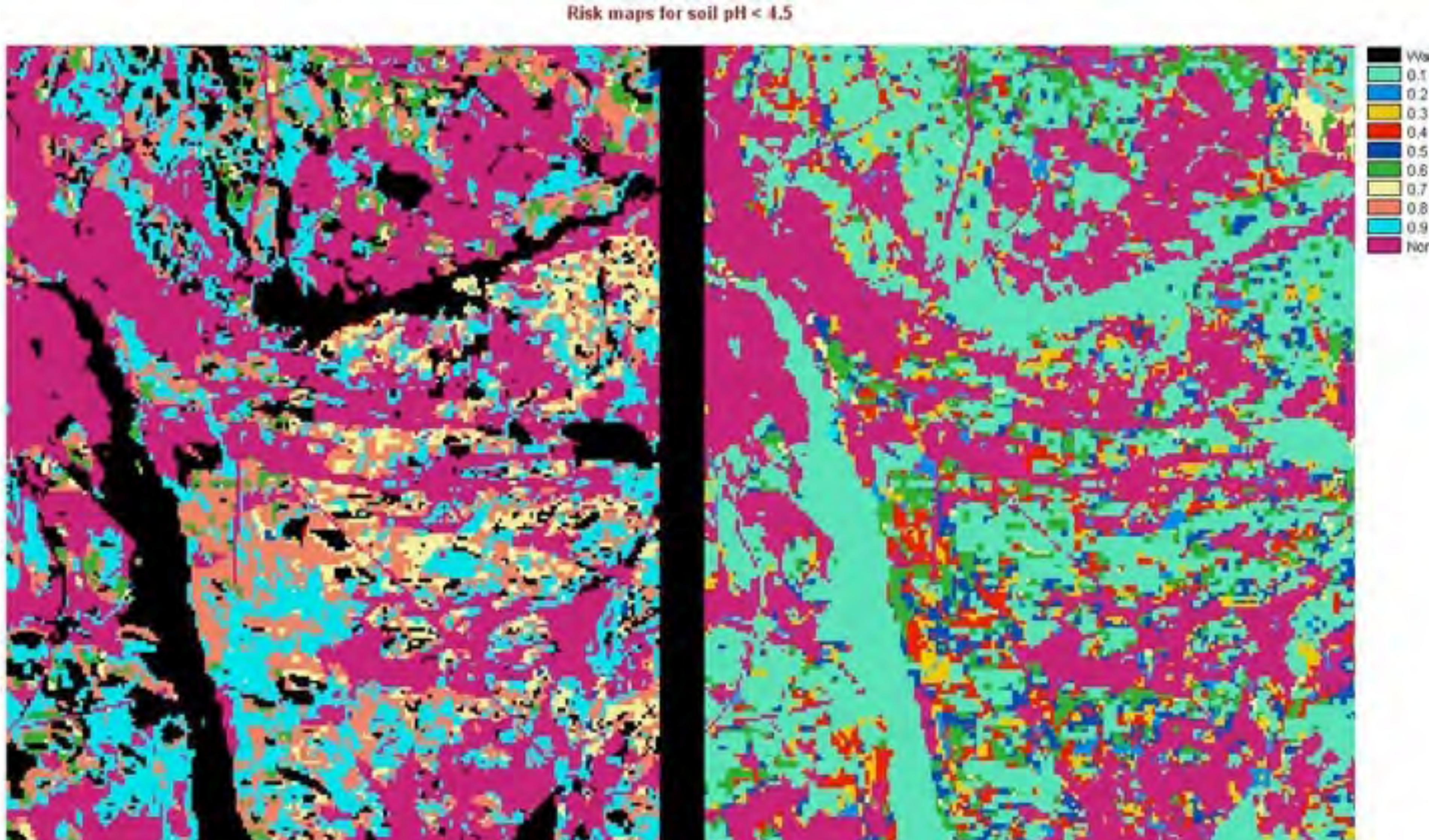
Crisp boundaries between soil polygons

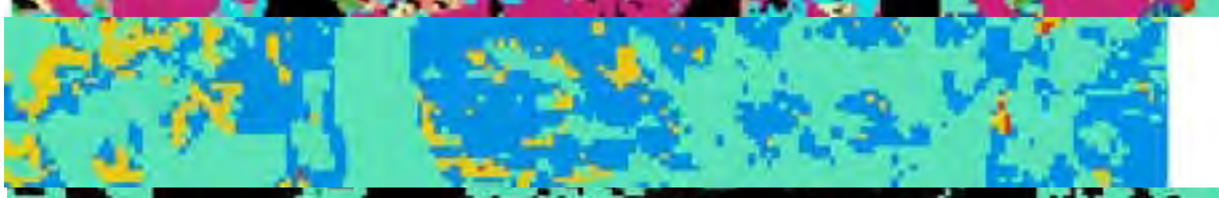
Fuzzy map



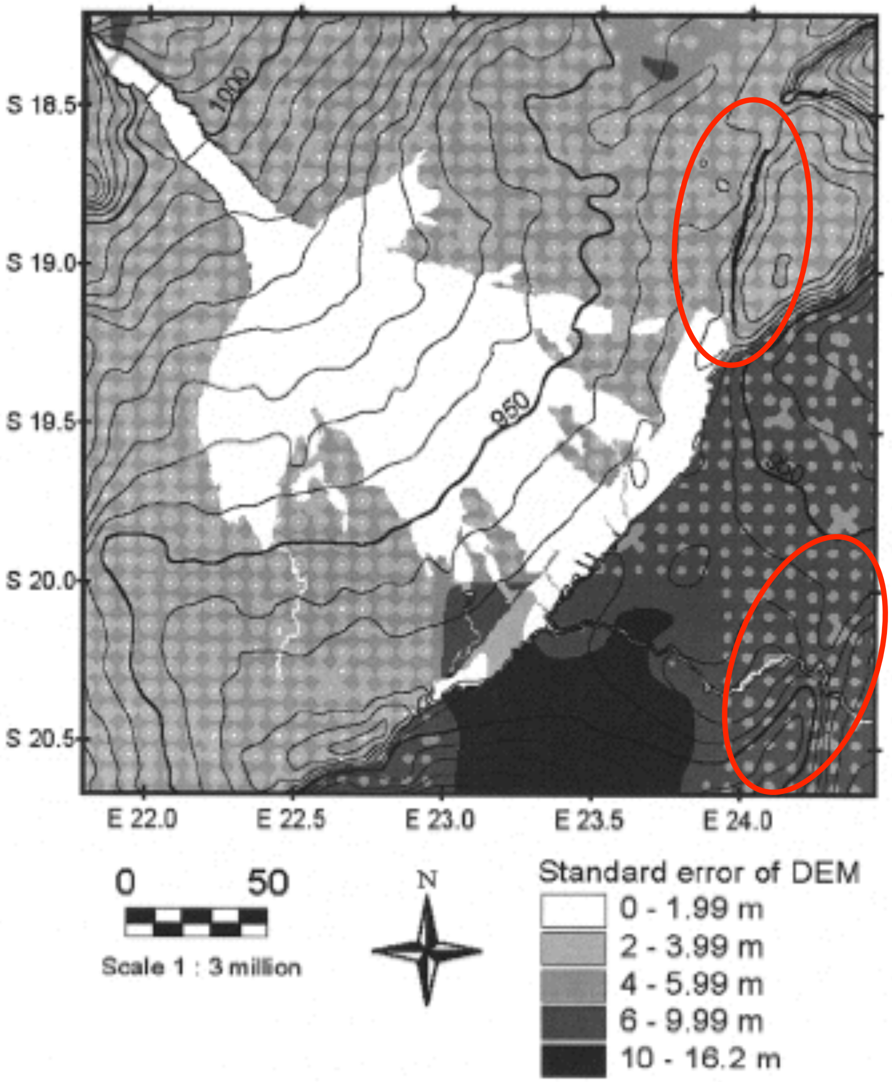
Imprecision bands around soil polygon borders

Osäkerhet vid mätning och modellering av pH värden





Okavango DEM



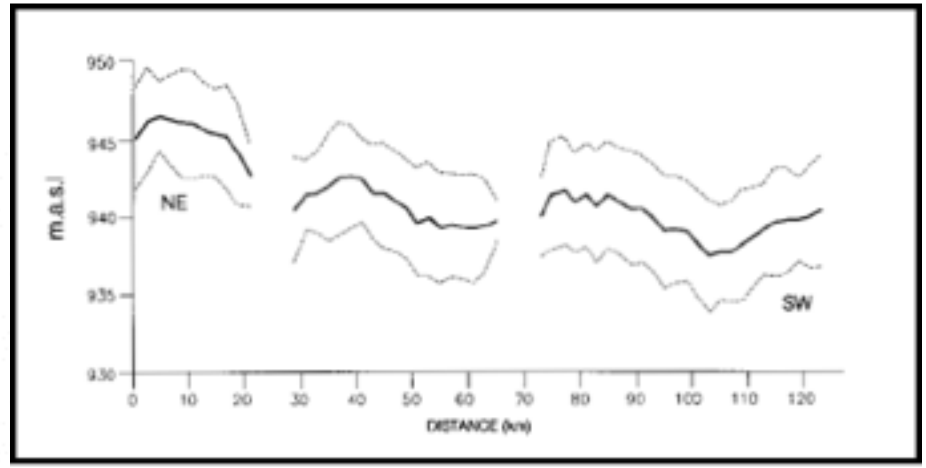


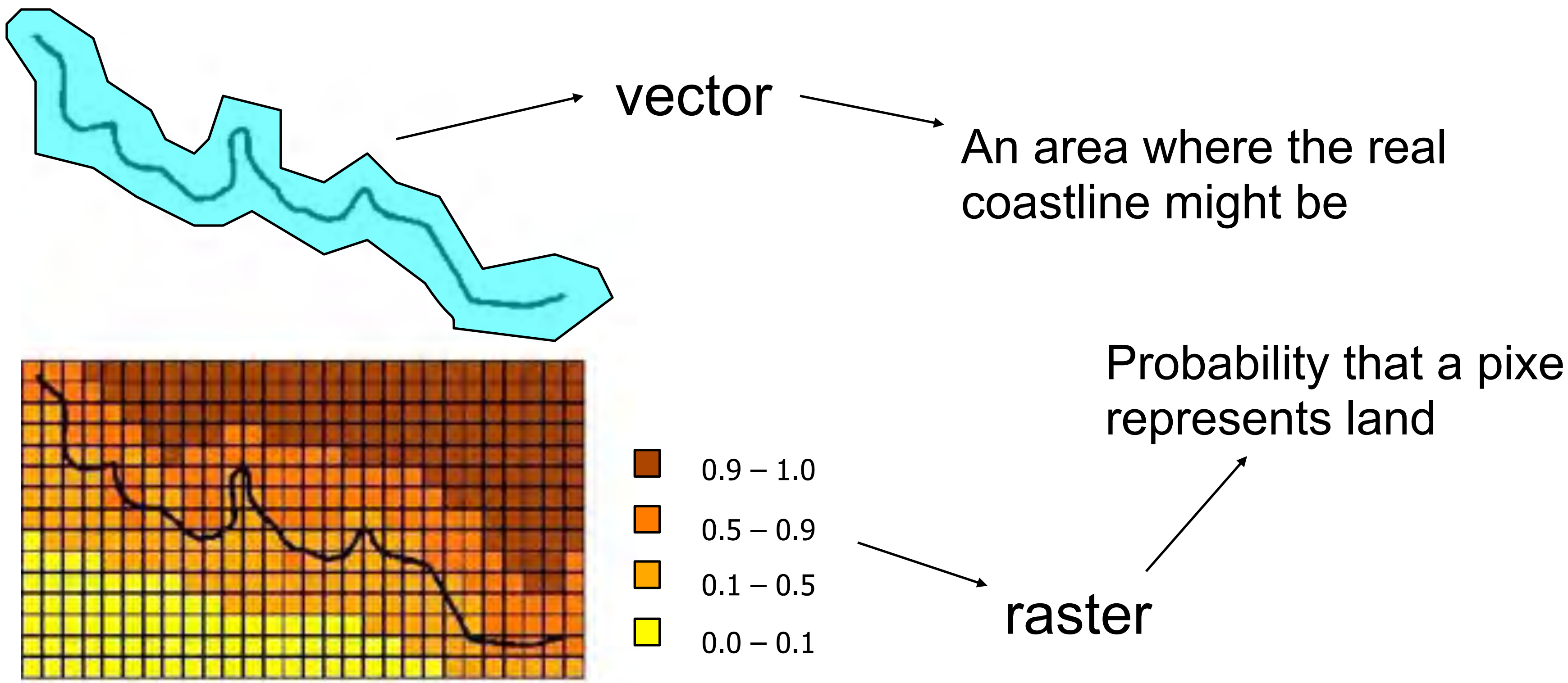
Figure 12b. Elevation profile along the western shoreline of palaeo-lake Makgadikgadi. The bold line shows the estimated elevation, while the two enveloping lines represent the standard deviation. Breaks occur in the lines where the palaeo-shoreline cannot be identified on the satellite image.

shoreline cannot be identified on the satellite image.

U2: Uncertainty in the measurement and representation

Representational models filter reality & uncertainty differently:

Uncertain generalised coastline



Example of an uncertain generalised coastline: fjords on a map of Scandinavia

Statistical models of uncertainty

For nominal attributes

How to measure the accuracy of nominal attributes?
e.g., a vegetation cover map

The confusion matrix

compares recorded classes (the observations) with classes obtained by some more accurate process, or from a more accurate source (the reference)

Reference (correct) classes

Observed (assigned) classes

	FAC	GSP	LOC	ORG	PER
FAC	100	0	0	0	0
GPE	0	98	0	0	2
LOC	0	5	95	0	0
ORG	1	1	0	92	6
PER	0	2	0	0	98

Example of a **misclassification or confusion matrix**. A grand total of 304 parcels have been checked. The rows of the table correspond to the land use class of each parcel as recorded in the database, and the columns to the class as recorded in the field. The numbers appearing on the principal diagonal of the table (from top left to bottom right) reflect correct classification.

Reference (correct) classes

Observed (assigned) classes

	A	B	C	D	E	Total
A	80	4	0	15	7	106
B	2	17	0	9	2	30
C	12	5	9	4	8	38
D	7	8	0	65	0	80
E	3	2	1	6	38	50
Total	104	36	10	99	55	304

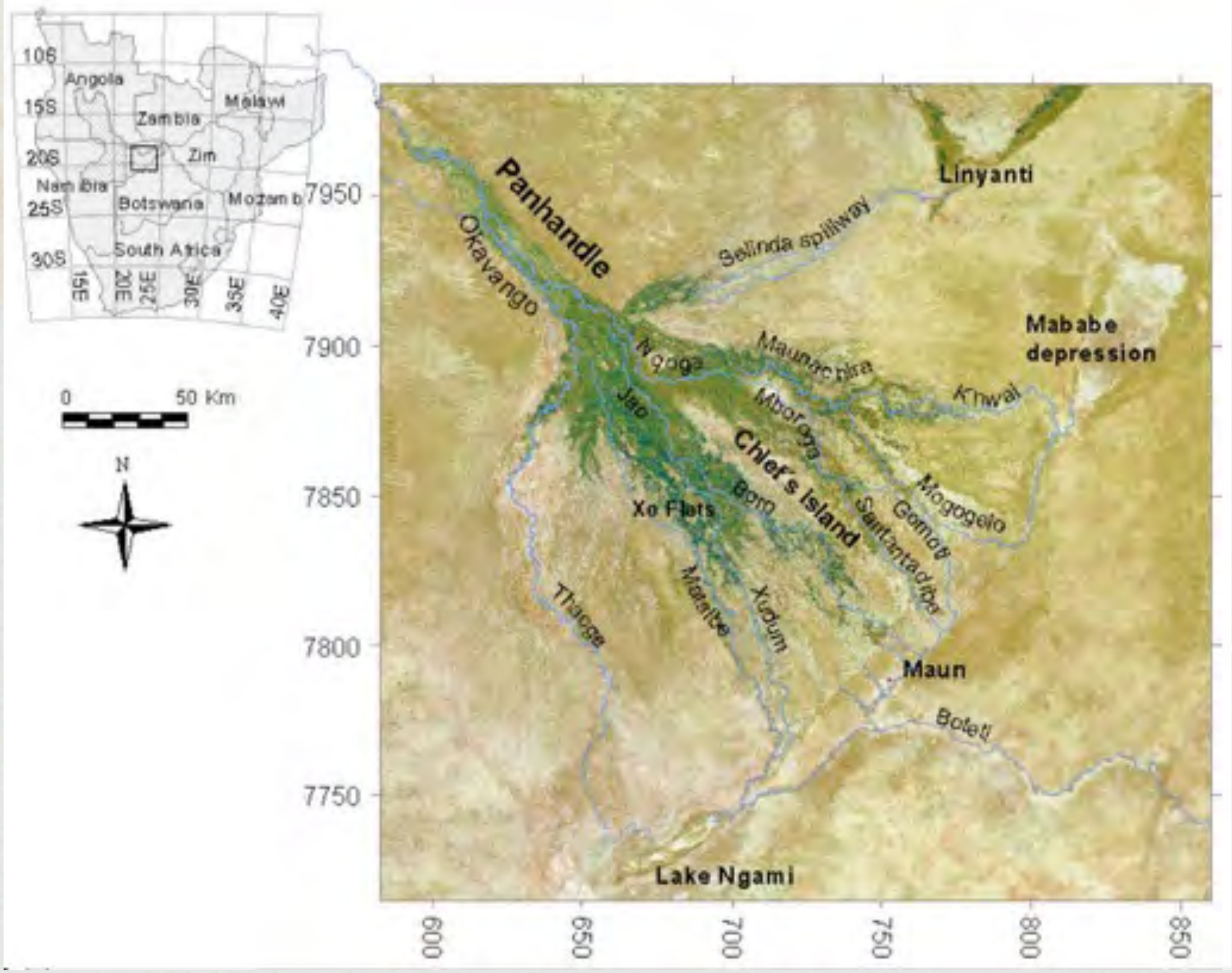
	A	B	C	D	E	Total
A	80	4	0	15	7	106
B	2	17	0	9	2	30
C	12	5	9	4	8	38
D	7	8	0	65	0	80
E	3	2	1	6	38	50
Total	104	36	10	99	55	304

How good is the classification?

Percent correctly classified parcels:
- total of diagonal entries divided by the grand total, times 100
- in this case: 209/304*100 = 68.8%

Kappa statistic = a special index to evaluate how good the assignment is:
- normalized to range from 0 (chance) to 100
- evaluates to 58.3% in this case

Okavango Delta - Botswana - Land cover classification



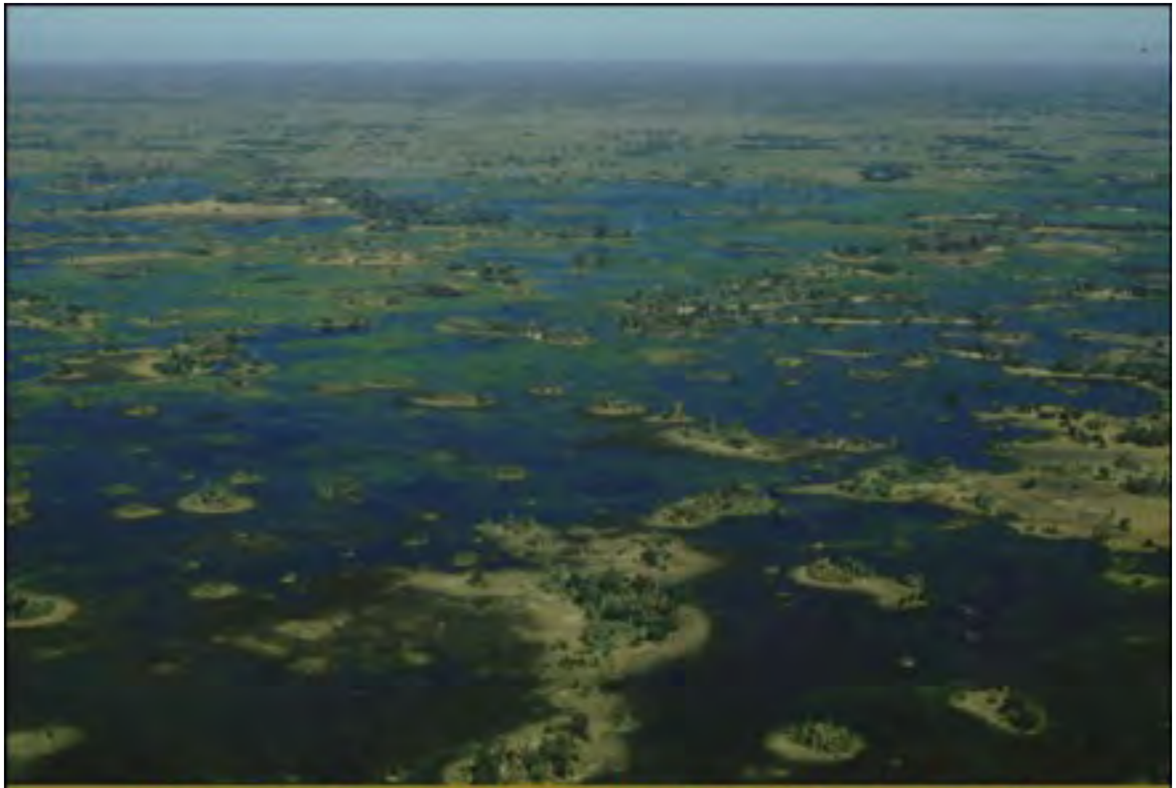
Water = 2.5 m below reference level

Permanent Swamp = 2.0 m below reference level

Permanent Swamp (Papyrus & Reed)

Primary floodplain = 1.5 m below reference level

Secondary floodplain = 1.0 m below reference level



Secondary Floodplain (in flood)

Grassland = reference level

Salt pan = 0.5 m below reference level

Dry Grassland/Salt Pan

Occasionally flooded grassland = 0.5 m below reference level

Salt pan = 0.5 m below reference level

Dry Grassland/Salt Pan (with flooding)

Riverine forest = 1.2 m above reference level

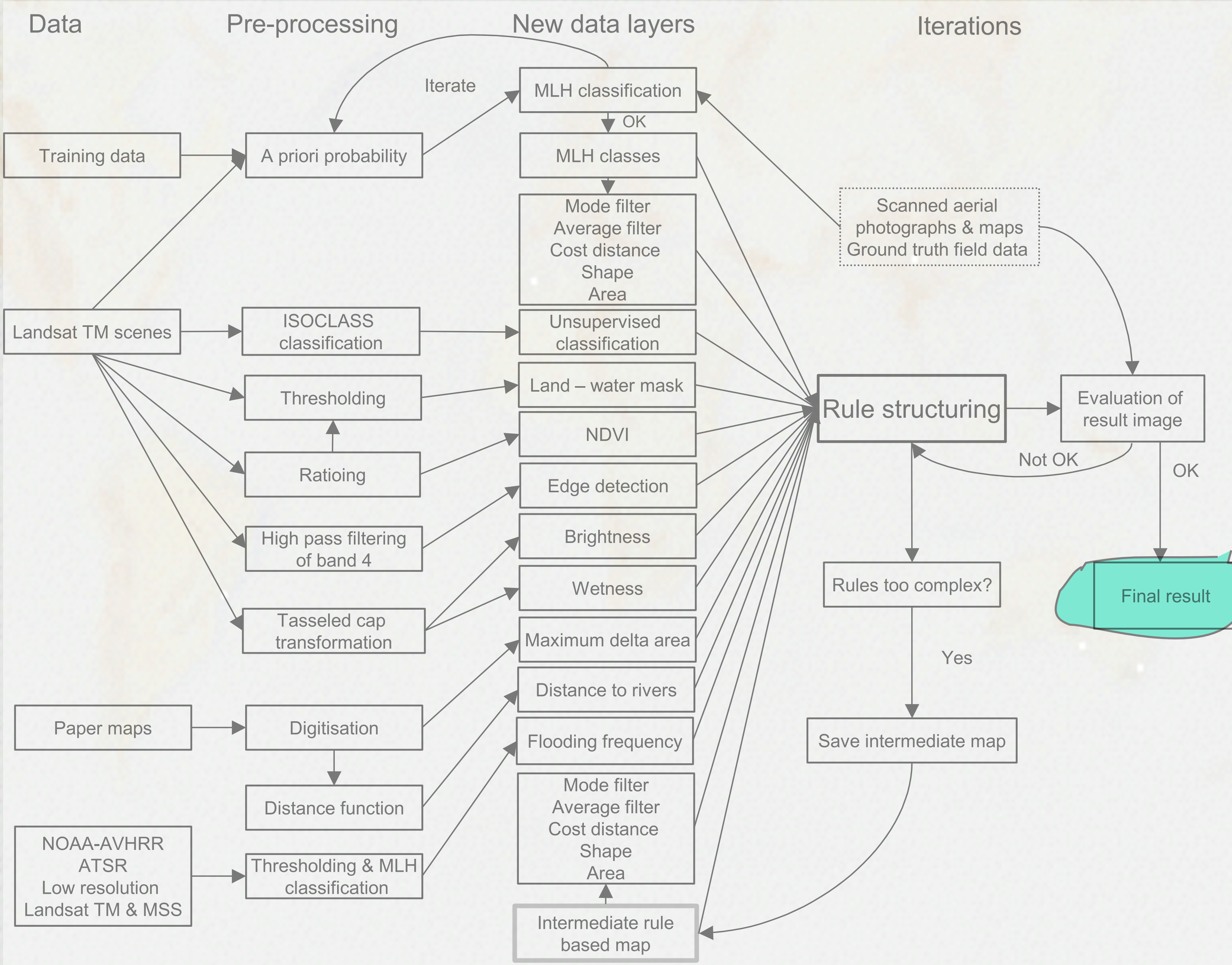
Dry woodland = reference level

Dry Woodland (dominated by Mopane)

Dry woodland = reference level

Dry Woodland (dominated by Acacia)

Okavango Delta knowledge based classification



Landcover ecoregions

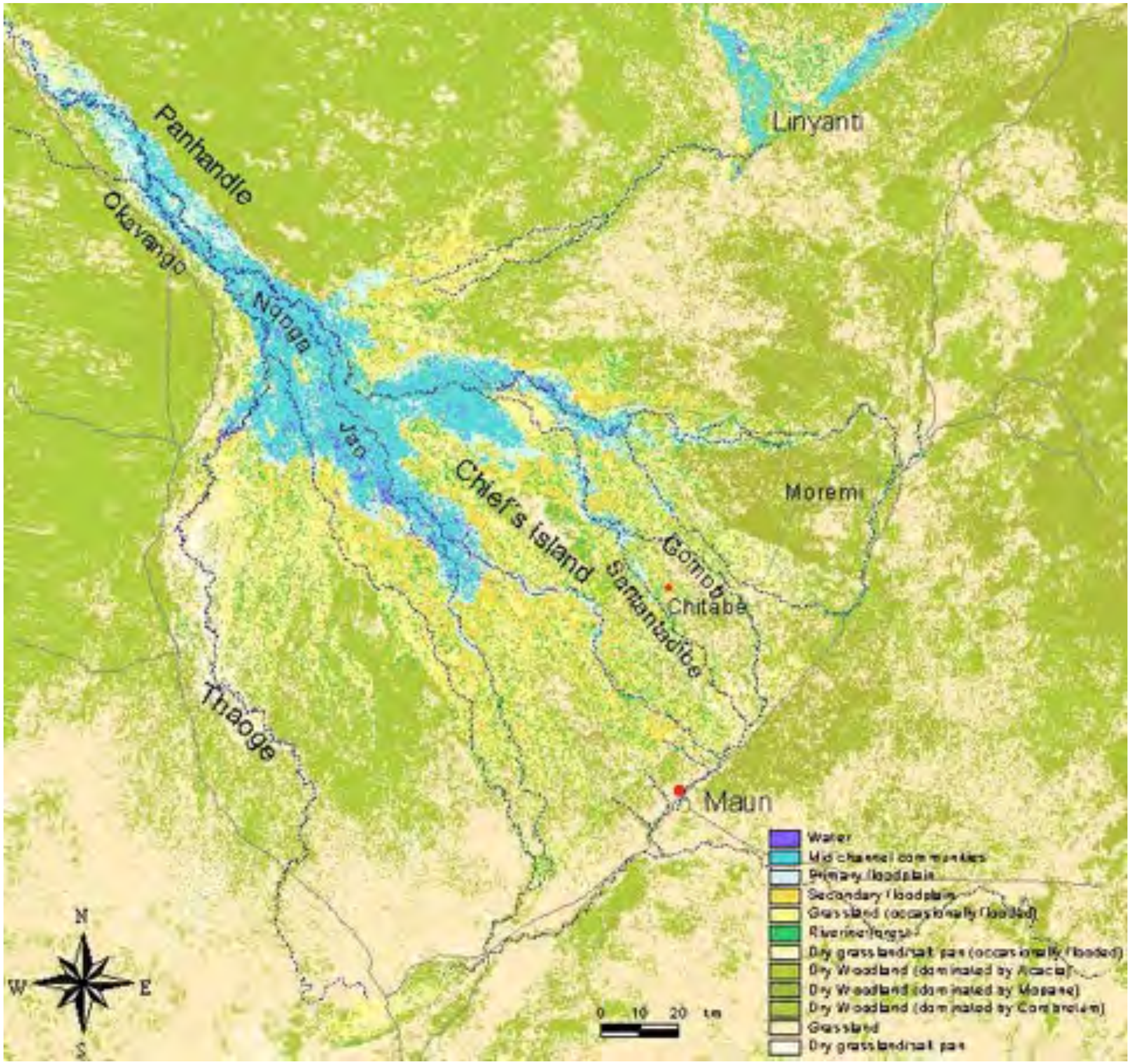


Table 5. Error matrix and accuracy for the rule-based classification in 10 classes (the columns represent the ground-data image and the rows the classified image).

	Wat	PSC	P FP	S FP	GL	SP	RF	Aca	Mop	Com	Total	ErrorC		
Wat	861	20	21	0	0	0	1	0	0	0	903	0.046	Wat	Water
PSC	15	1000	234	21	0	5	106	1	0	0	1382	0.276	PSC	Permanent swamp communities
P FP	9	135	980	279	27	0	160	2	0	3	1595	0.386	P FP	Primary flood plain
S FP	1	14	57	149	2	0	78	8	0	1	310	0.519	S FP	Secondary flood plain
GL	0	0	97	76	191	11	46	328	56	2	807	0.763	GL	Grassland
SP	0	0	0	0	71	98	2	73	15	2	261	0.624	SP	Sparse grassland/salt crust
RF	4	7	13	5	4	3	382	190	0	20	628	0.392	RF	Riparian forest
Aca	0	0	0	2	0	4	34	362	57	3	462	0.216	Aca	Acacia woodland
Mop	0	0	0	0	0	0	14	79	106	0	199	0.467	Mop	Mopane woodland
Com	0	0	0	0	0	0	3	2	1	0	6	1.000	Com	Combretum woodland
Total	890	1176	1402	532	295	121	826	1045	235	31	6553		ErrorC	Error of commission
ErrorO	0.033	0.150	0.301	0.720	0.352	0.190	0.538	0.654	0.549	1.000		0.370	ErrorO	Error of omission

Table 6. Error matrix and accuracy for the rule-based method in six classes (the columns represent the ground-data image and the rows the classified image).

	Wat	PSC	FP	GL	SP	For	Total	ErrorC		
Wat	861	20	21	0	0	1	903	0.046	Wat	Water
PSC	15	1000	255	0	5	107	1382	0.276	PSC	Permanent swamp communities
FP	10	149	1465	29	0	252	1935	0.231	FP	Flood plain
GL	0	0	173	191	11	432	807	0.763	GL	Grassland
SP	0	0	0	71	98	92	261	0.624	SP	Sparse grassland/salt crust
For	4	7	20	4	7	1253	1295	0.032	For	Forest
Total	890	1176	1934	295	121	2137	6553		ErrorC	Error of commission
ErrorO	0.033	0.150	0.242	0.352	0.190	0.414		0.257	ErrorO	Error of omission

The abbreviations correspond to the six aggregated ecoregion classes. Overall accuracy = 74.3%; Kappa index of agreement = 0.67.

Fuzzy logic expert classifier used for the Okavango landcover

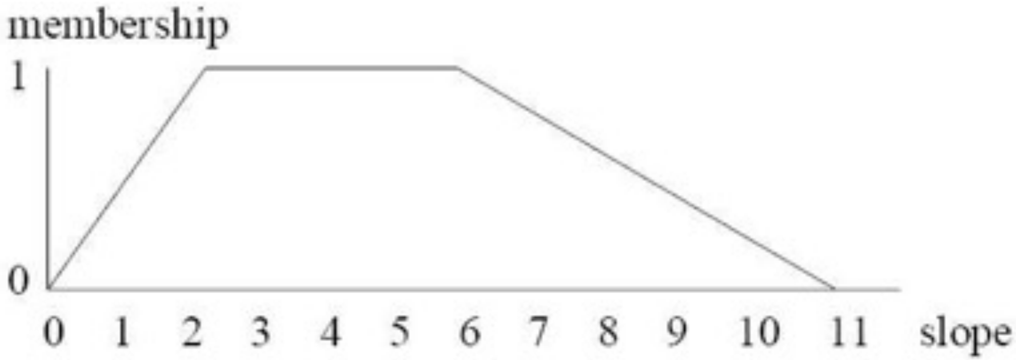
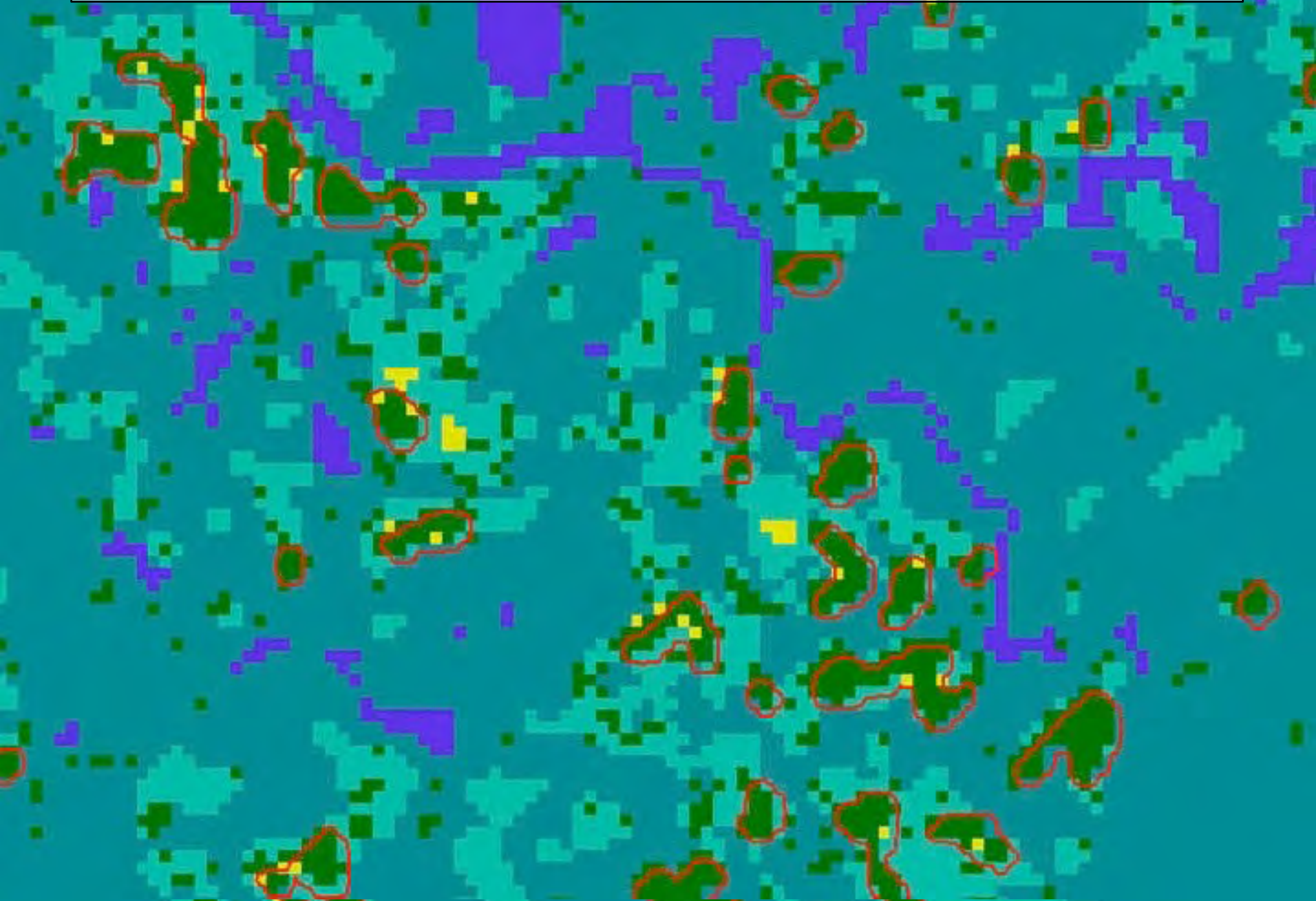


Figure 3. Example of fuzzy membership function to slope. Between slopes 2 and 6 the membership is 1, whereas it changes gradually from 0 to 1 between slope 0 and 2 and 11 and 6 respectively. In guide the above fuzzy membership function is written "Whenimg @ 0 2 TO 6 11 Slope", the Boolean logic is written "Whenimg @ 2 TO 6 Slope".

Table 1. Command structure in guide *(forest in the example can either be classified as a lumped category, or the two observations can be separately classified, making it possible to include several forest types in one rule).*

command	Followed by	Example - Boolean logic	Example - fuzzy logic
WHENIMG	= category "map"	= 3 4 soil	= 3 4 soil
	< category "map"	> 0 300 DEM	>0 40 300 500 DEM
	> category "map"	< 30 20 slope	< 50 30 30 20 slope
	@ categ TO categ "map"	@ 10 TO 40 20 TO 30 LAI	@ 0 10 TO 40 50 10 20 TO 30 40 LAI
	+ row nr TO row nr	+ 0 TO 150 100 TO 250	+ 0 0 TO 150 300 50 100 TO 250 300
	* column nr TO col nr	* 100 TO 500 400 TO 600	* 50 100 TO 500 600 200 400 TO 600 700
SAVEIMG	# category "name"	# 5 forest	# 5 forest

Transformations fel vid varpning av raster jämfört med vektor



Statistical models of uncertainty

For interval/ratio attributes

Measurements of attribute values are distorted by small amounts = **errors**

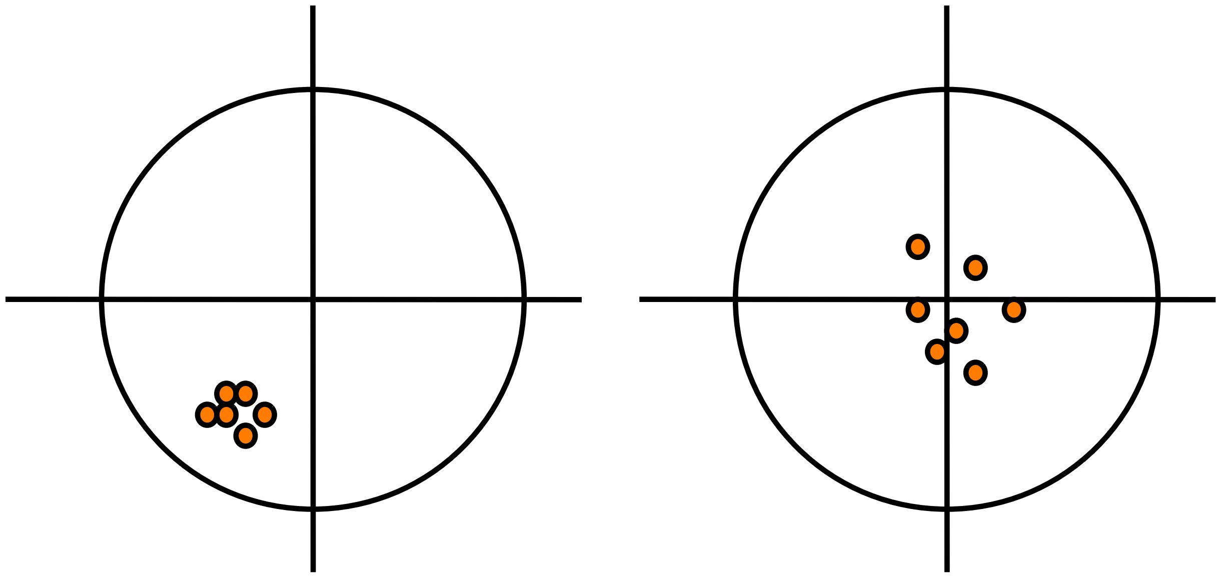
Uncertainty expressed with two concepts

Precision

Accuracy = the amount of distortion from the true value

the variation among repeated measurements

also the amount of detail in the reporting of a measurement



The term *precision* is often used to refer to the repeatability of measurements. In both diagrams six measurements have been taken of the same position, represented by the center of the circle. On the left, successive measurements have similar values (they are *precise*), but show a bias away from the correct value (they are *inaccurate*). On the right, precision is lower but accuracy is higher.

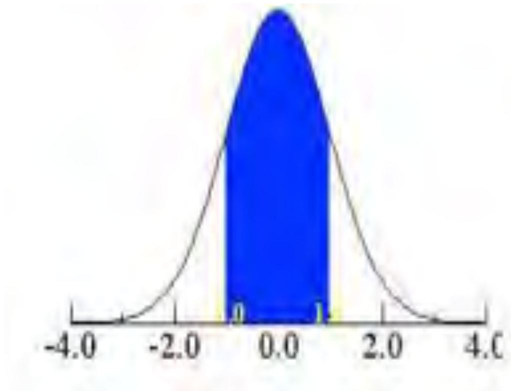
The primary measure of accuracy in map accuracy standards and GIS databases:

RMSE = Root Mean Square Error =
the square root of the average squared error

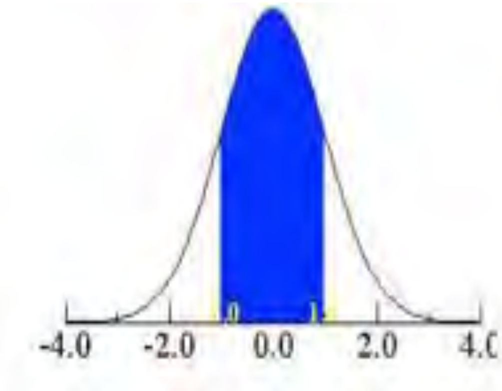
For practical purposes: approximately equal to the absolute value of the average error in each observation

Additional measure of accuracy:

How errors are **distributed in magnitude**?
How many are small, how many are large?

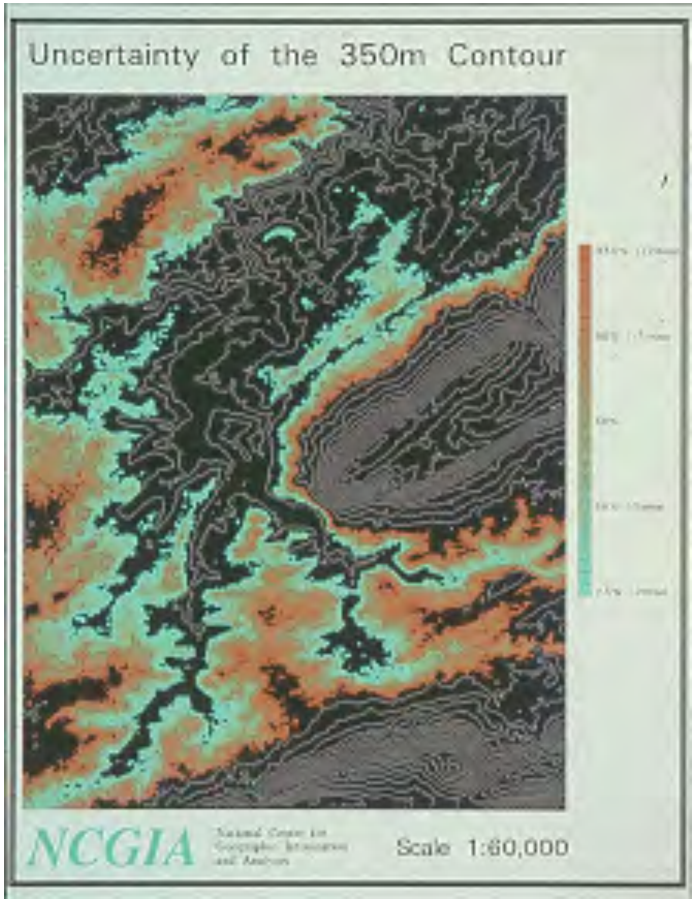


Unbiased measurements:
errors follow a normal distribution (Gaussian curve), with mean=0 (there is approximately the same amount positive and negative errors – positive and negative errors cancel each other out)



The Gaussian or Normal distribution. The height of the curve at any value of *x* gives the relative number of observations with that value of *x*. The area under the curve between any two values of *x* gives the probability that observations will fall in that range. The range between –1 standard deviation and +1 standard deviation is in blue. It encloses 68% of the area under the curve, indicating that 68% of observations will fall between these limits.

Uncertainty in the location of the 350 m contour based on an assumed RMSE of 7 m. The Gaussian distribution with a mean of 350 m and a standard deviation of 7 m gives a 95% probability that the true location of the 350 m contour lies in the colored area, and a 5% probability that it lies outside.



U3: Uncertainty in the analysis

Uncertainties in data → Uncertainties in the results of the analysis

Almost every input to a GIS is subject to error and uncertainty.

That is not always the case in practice!

In principle, every output should have confidence limits or some other expression of uncertainty.

Soil map

Soil map + uncertainty of borders

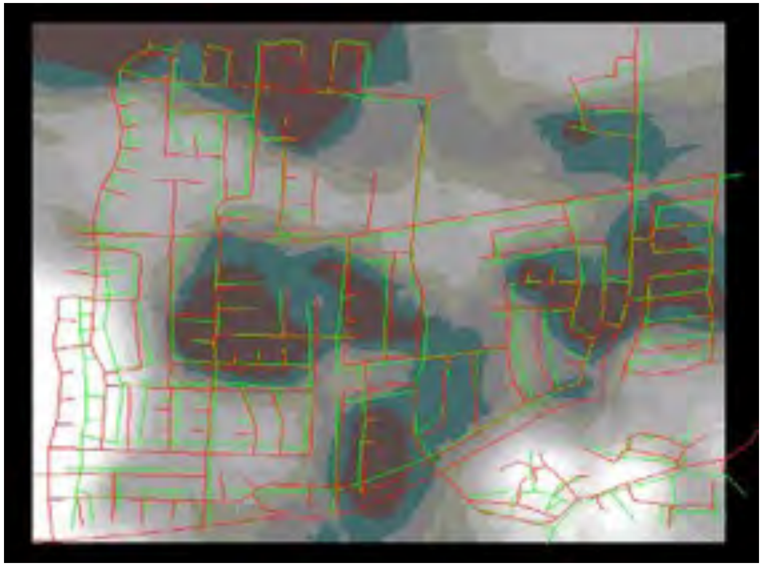
Validation → The process of confirming the validity of data, information or processes.

checking if something satisfies a certain criterion

How valid are results of a GIS analysis?

external

internal



estimating effect of uncertainty through modelling different possible outcomes

simulation

error propagation

combining&comparing data from different sources

Dealing with uncertainty

It is easy to see the importance of uncertainty in GIS, but much more difficult to deal with it effectively.

Basic principles for dealing with uncertainty:

- Uncertainty is inevitable in GIS.

- Data obtained from others should never be taken as truth.
Instead, efforts should be made to determine quality.

- Effects on GIS outputs are often much greater than expected.
This is important, as people tend to regard computer outputs as the truth.

- Use as many sources of data as possible.
Cross-check them for accuracy

- Be honest and informative in reporting analysis results.
Add warnings and cautions.

Final project – E8

Task: solve GIS problems similar to the ones you have solved during the exercises

Find a solution to a proposed problem

Submit a correctly written report

Present the project to the class

Work in **own time.**

Faggott & MacLean

GISlab

Help is available during scheduled times.

Work should be done in groups of 2-3 students (same as for all the exercises). All students in each group are required to take part in the presentation.

This is a small project! → Do not use more time than that!

→ Equivalent to two 4-hours labs.

Project topics

→ Select one from the list of proposals

→ Suggest your own topic (find your own problem & data)

→ Modify one of the proposals

However, to keep things from being too repetitive during presentations:
- maximum 3 groups per each topic proposal
- topics allocated on a first-come first-serve basis

→ Decide what you want to do and send me an email with your selected topic.